计算系统生物学

王 勇

中国科学院数学与系统科学研究院

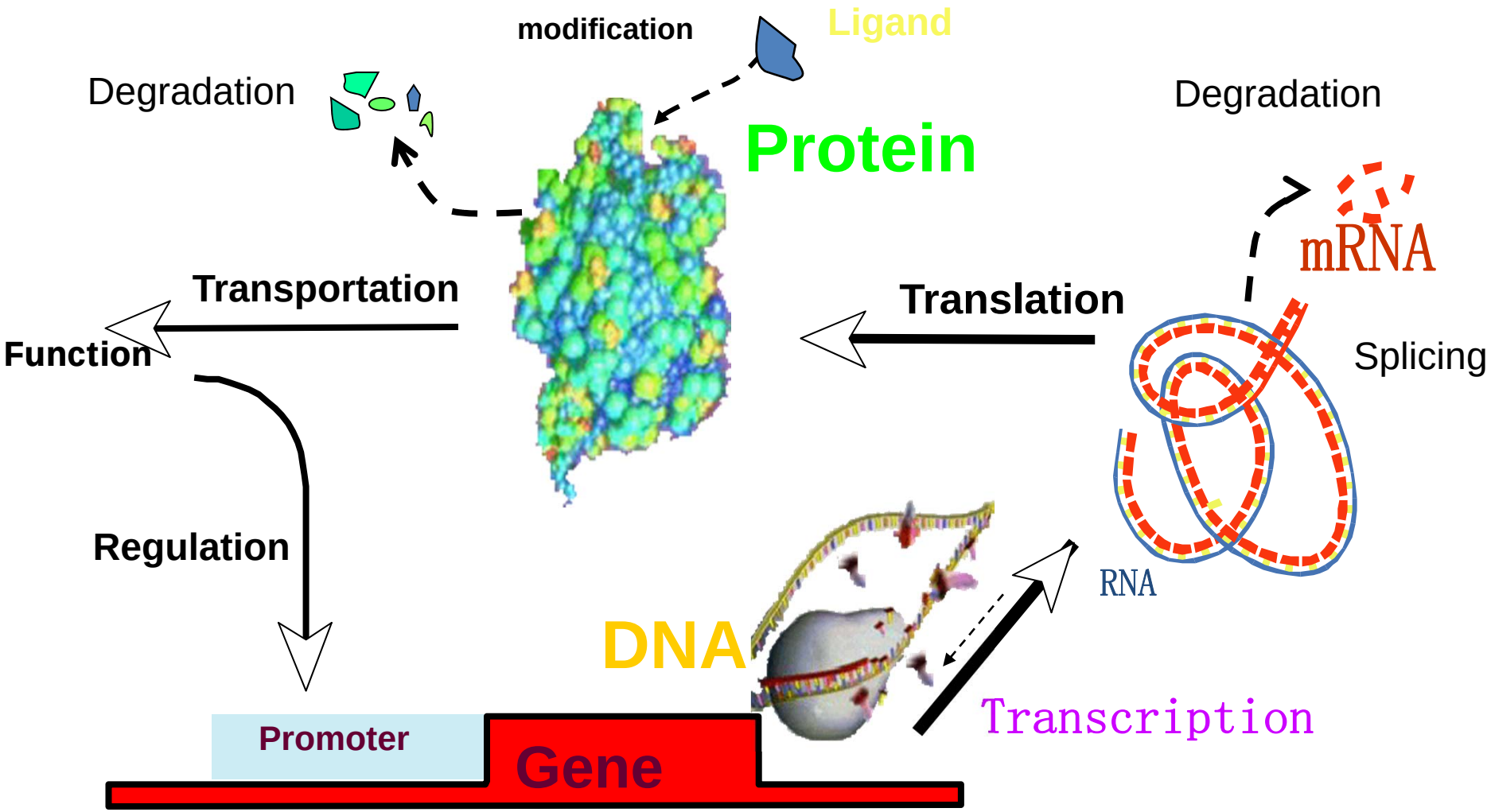
Gene Regulatory Network Inference

In Systems Biology Framework

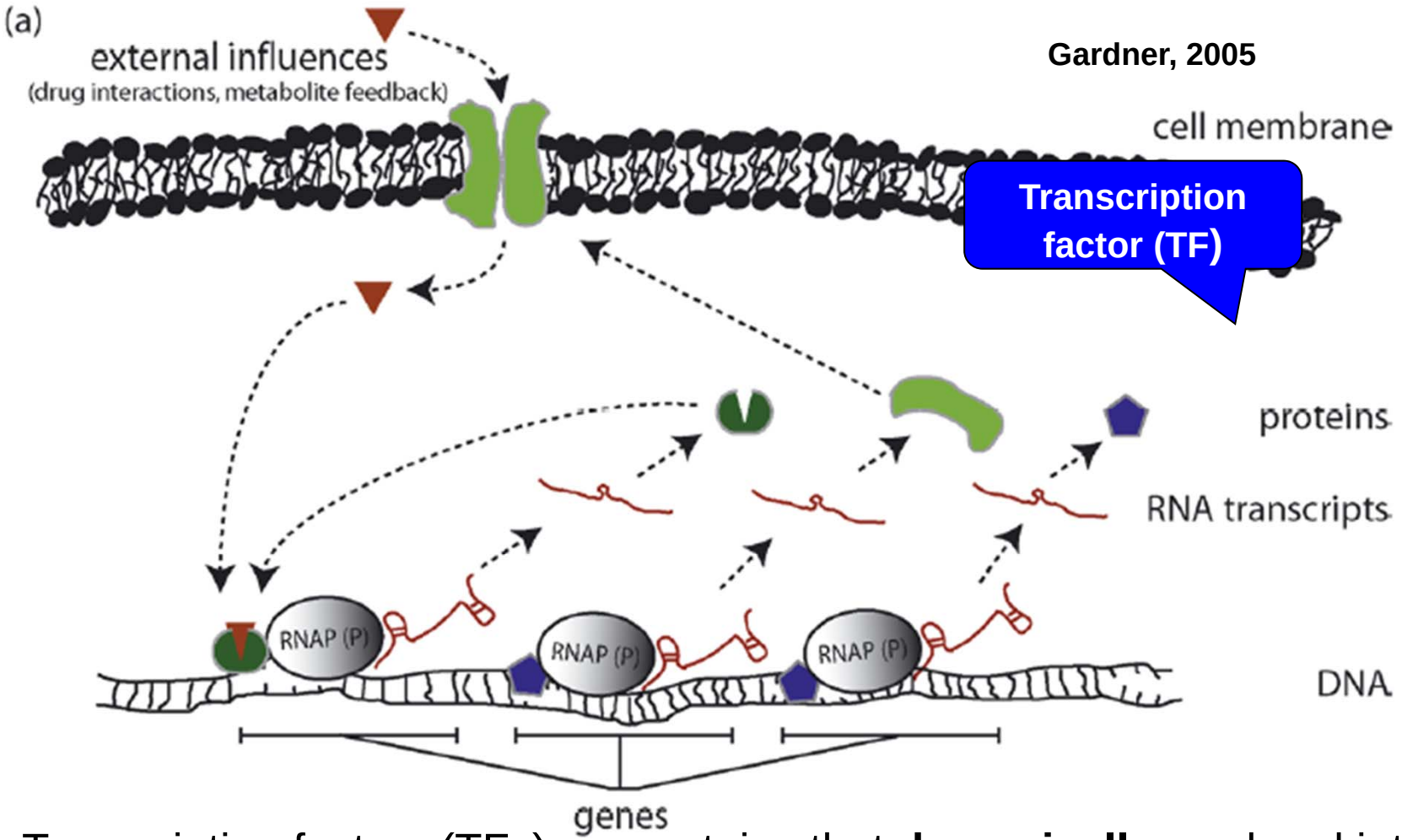
Yong Wang

http://zhangroup.aporc.org

Central dogma of molecular biology

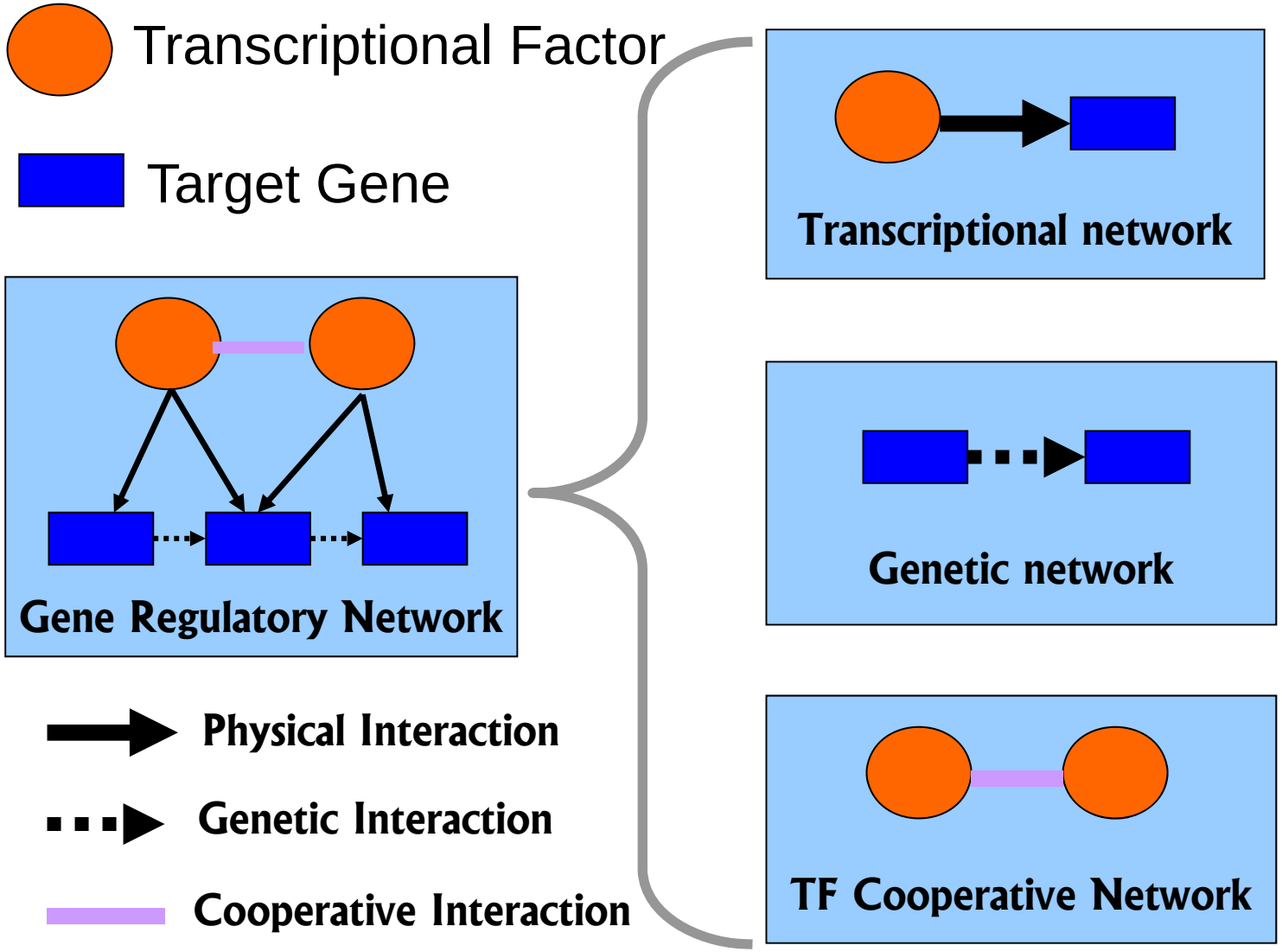


Gene regulation

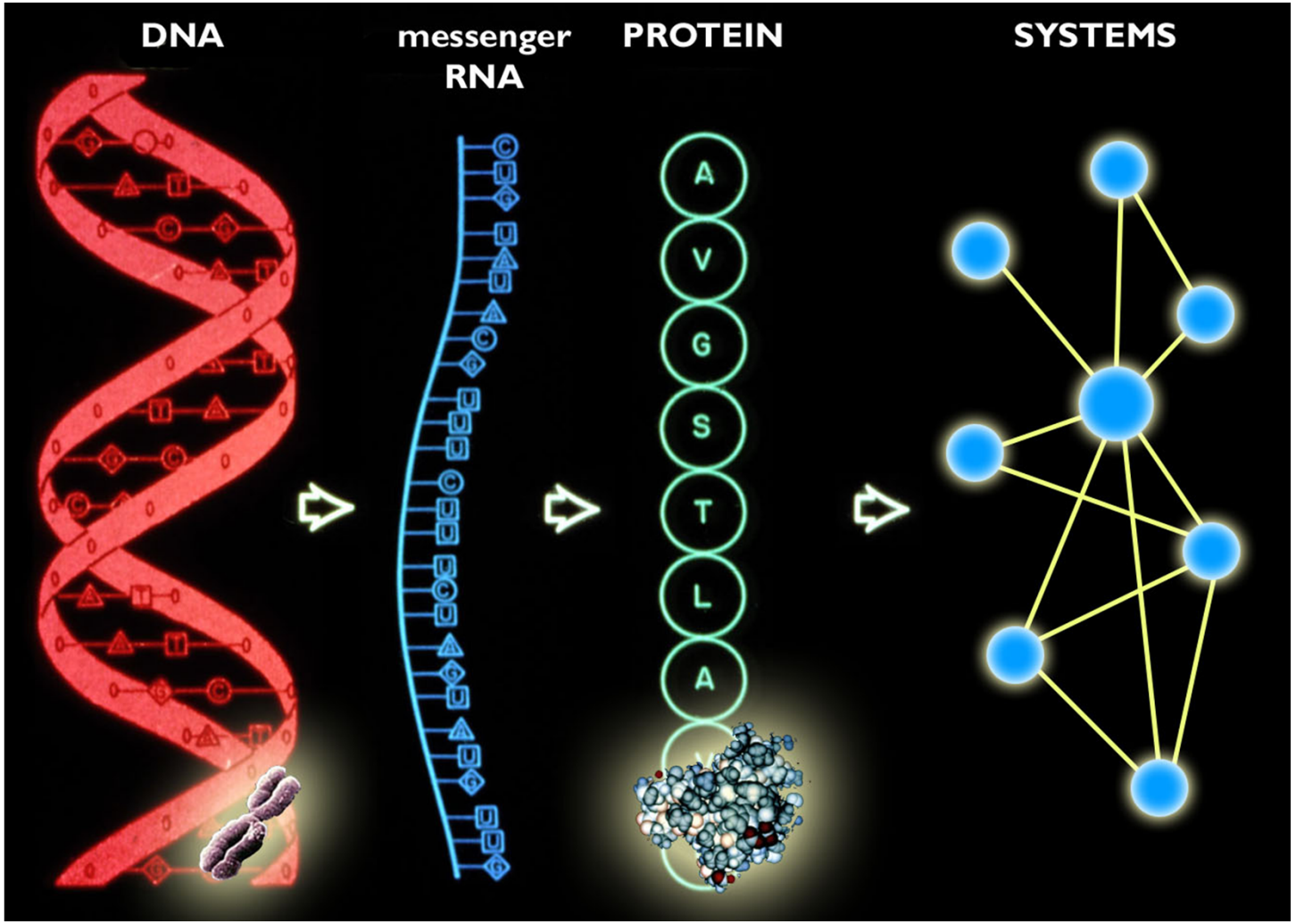


Transcription factors (TFs) are proteins that **dynamically** read and interpret the **static** genetic instructions in the DNA

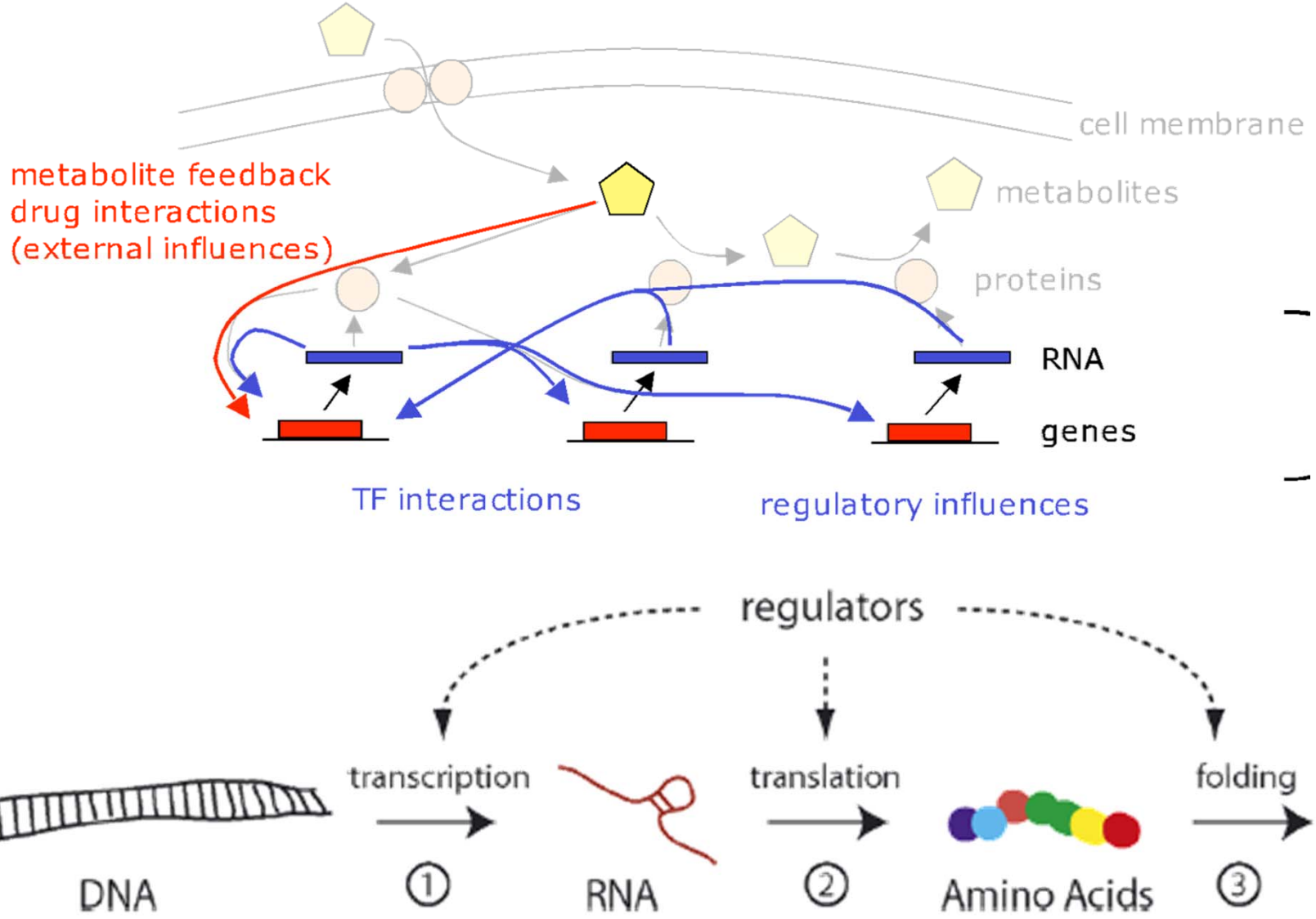
Basic building blocks for gene regulatory network

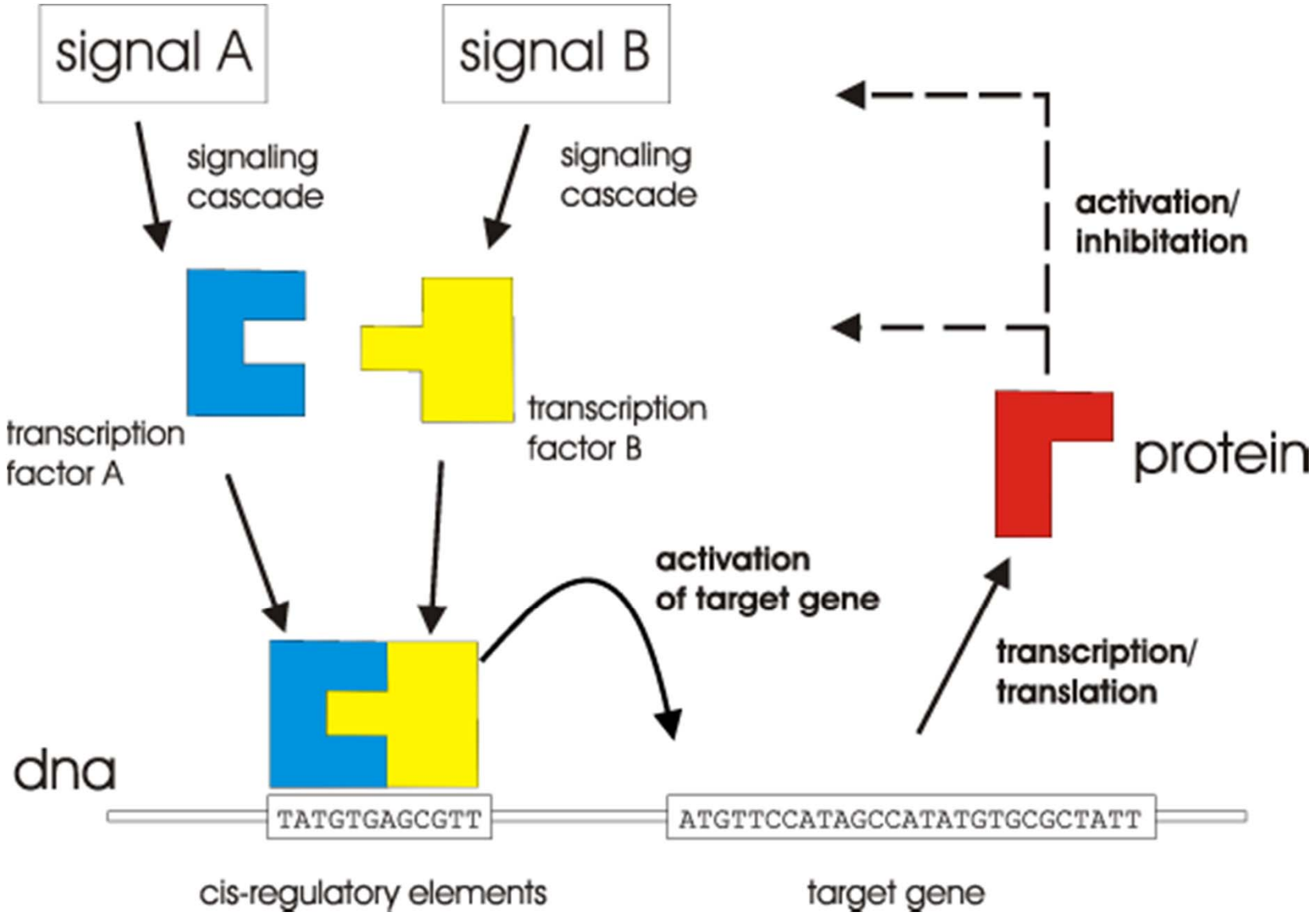


Background---GRN

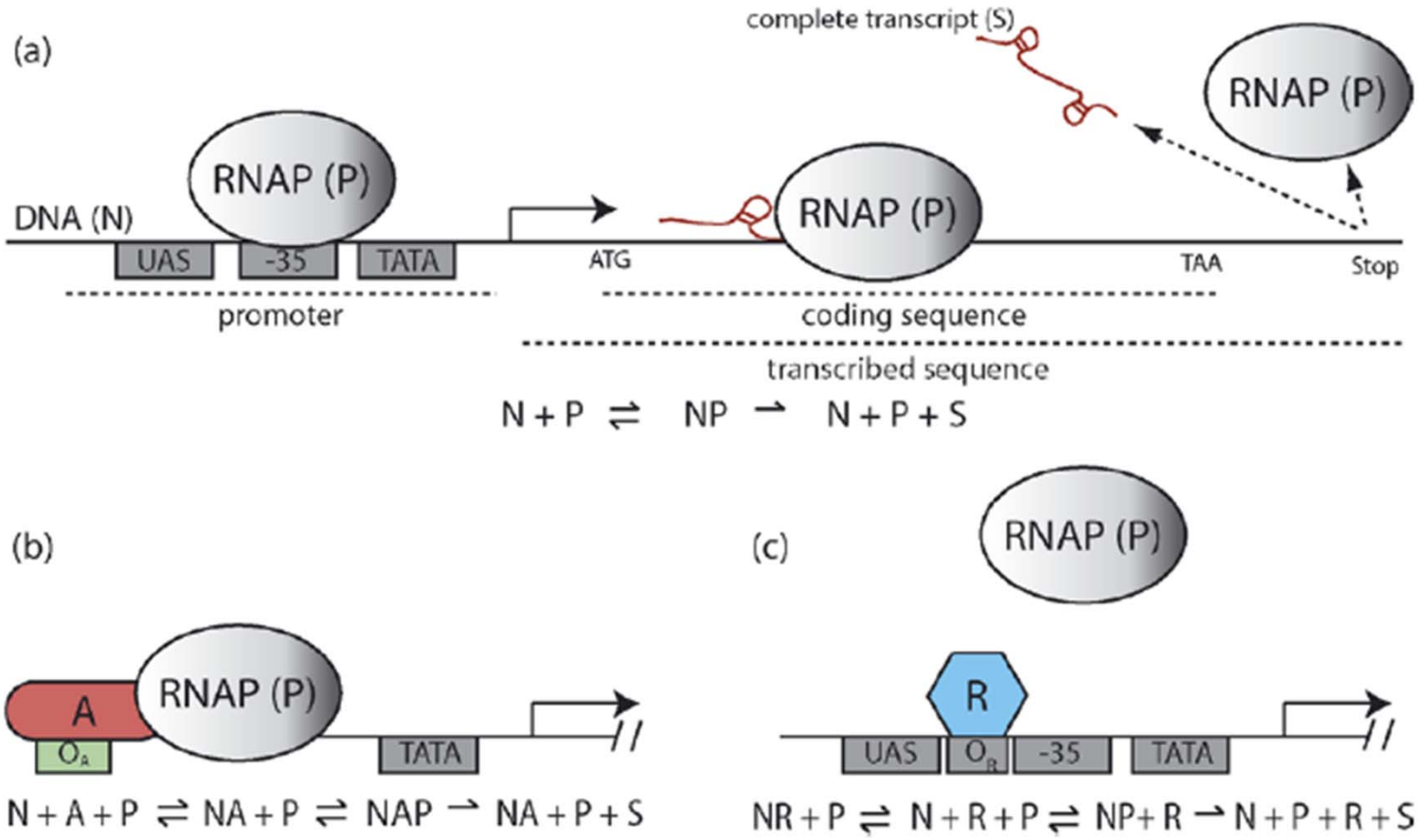


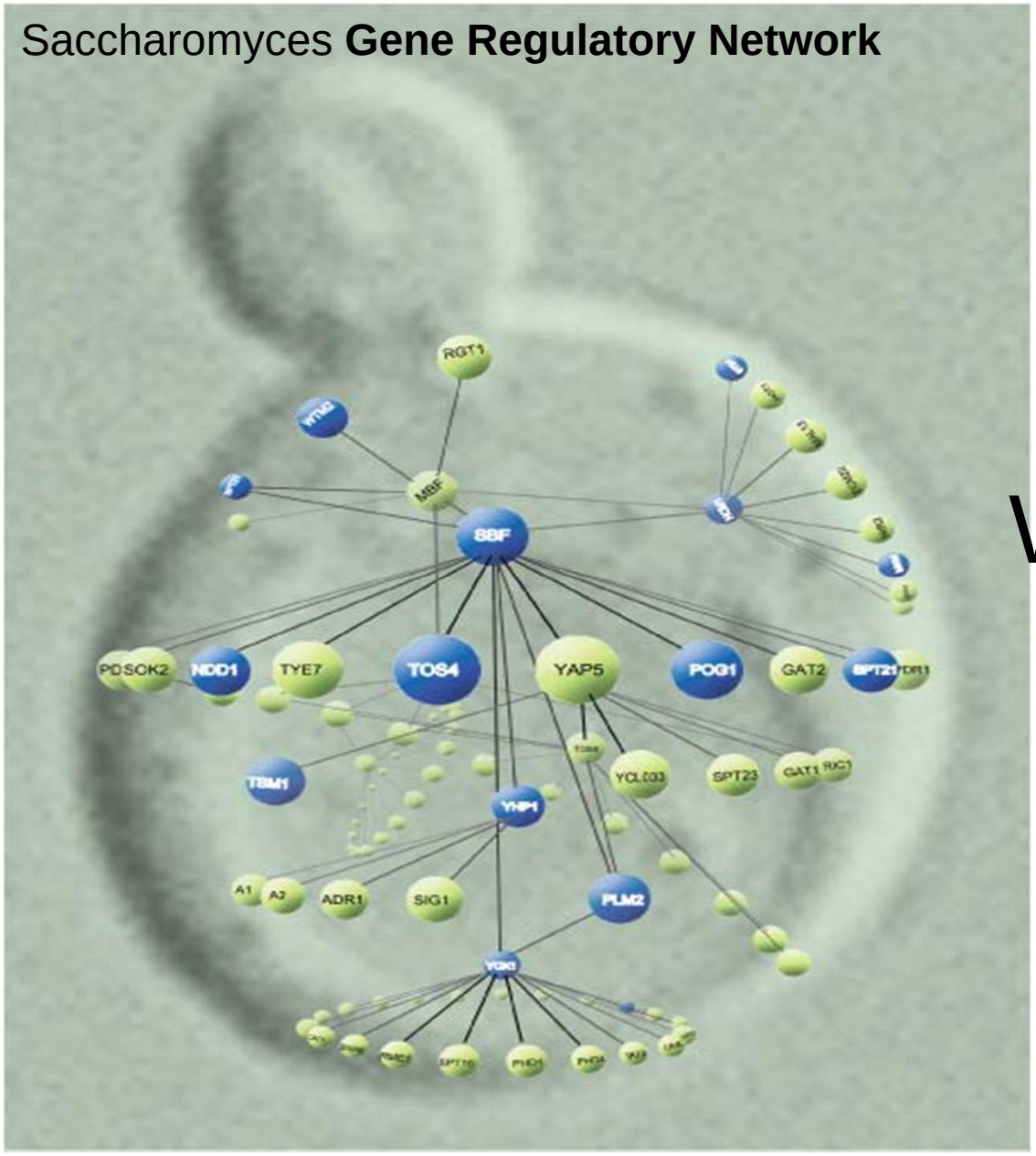
Biological GRN




signal A
signal B
signaling cascade
signaling cascade
activation/ inhibitation
transcription factor A
transcription factor B
protein
activation of target gene
transcription/ translation
dna
TATGTGAGCGTT
ATGTTCCATAGCCATATGTGCGCTATT
cis-regulatory elements
target gene

Activation or repression

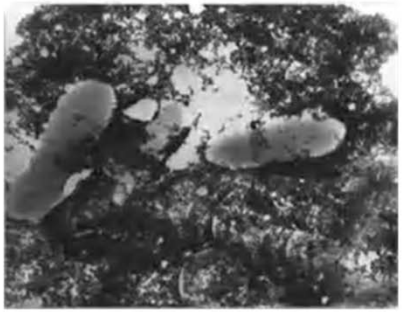



Saccharomyces **Gene Regulatory Network**
RGT1
MBF
SBF
PDSOK2
NDD1
TYE7
TOS4
YAP5
POG1
GAT2
TBS1
YCL033
SPT23
YHP1
A1
A2
ADR1
SIG1
PLM2

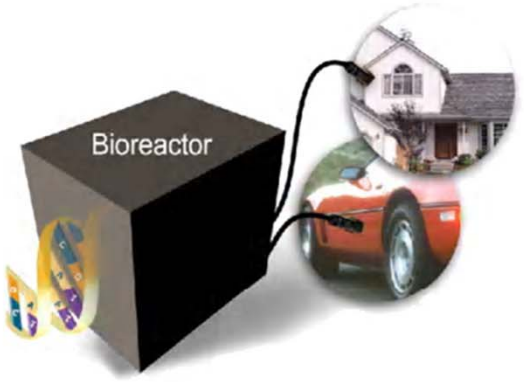
What we want?

Network Inference, Analysis and Control

Drug Discovery

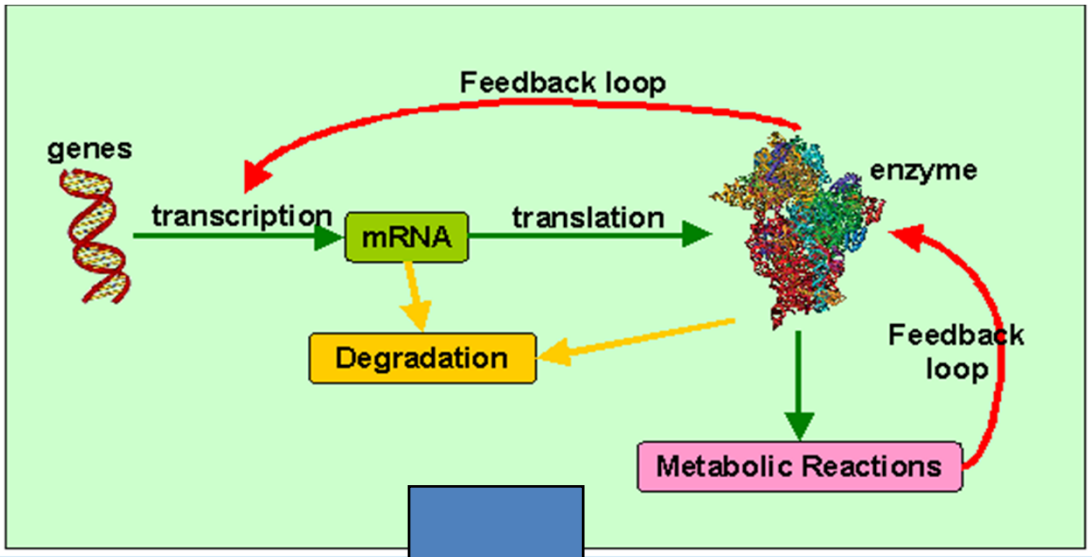


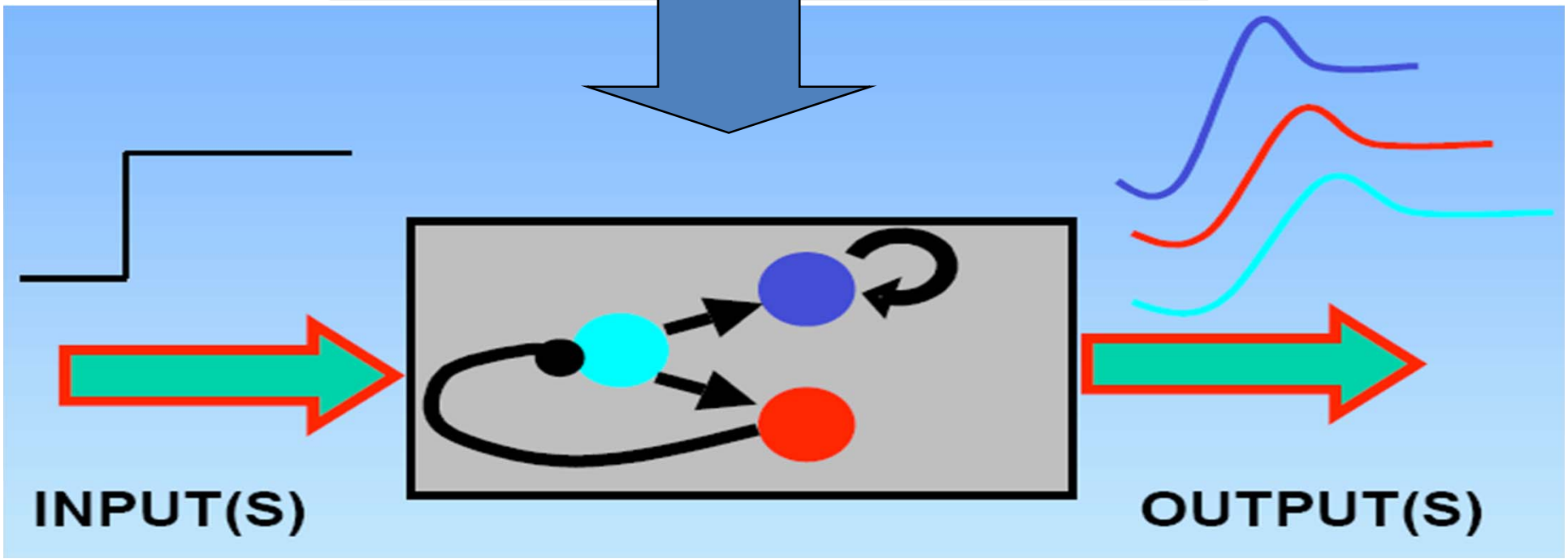
Bioenergy



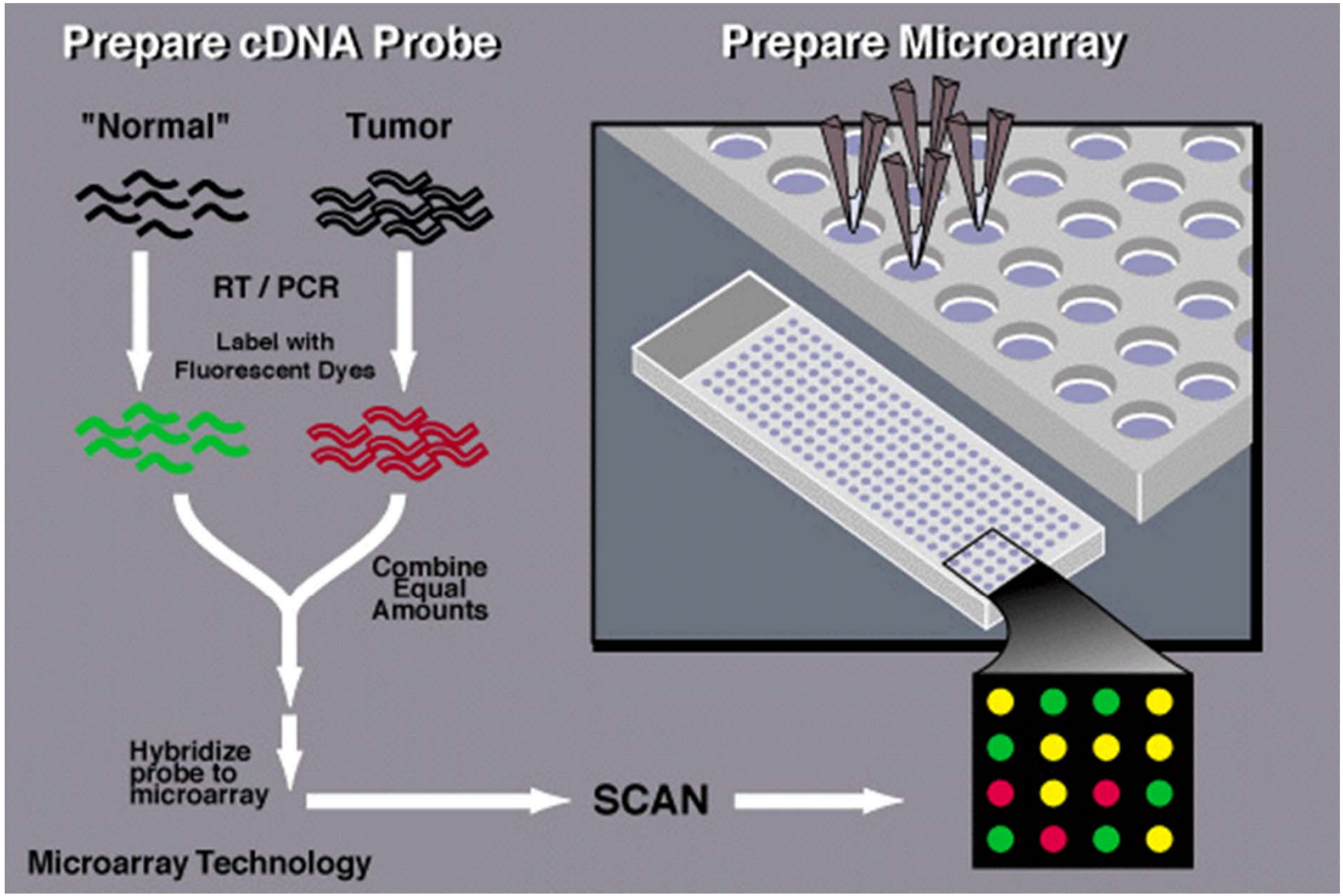
Bioremediation

How to?

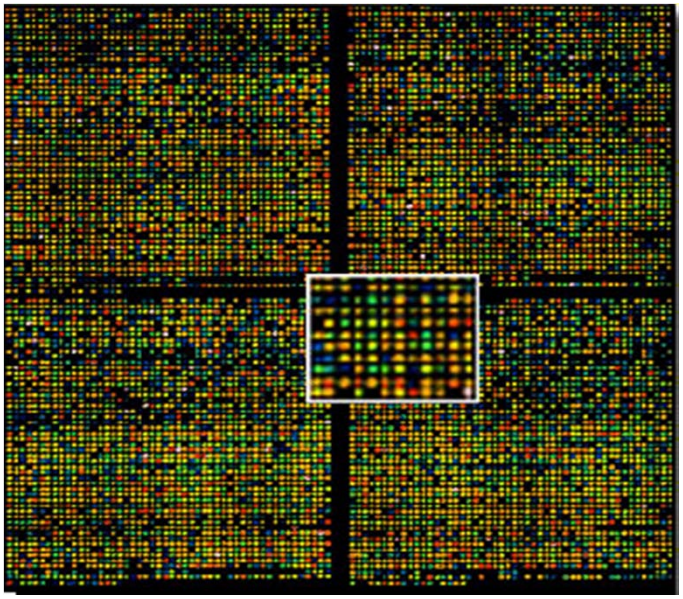

Feedback loop
genes
transcription
mRNA
translation
enzyme
Degradation
Feedback loop
Metabolic Reactions


INPUT(S)
OUTPUT(S)

Feasibility: Microarray technology



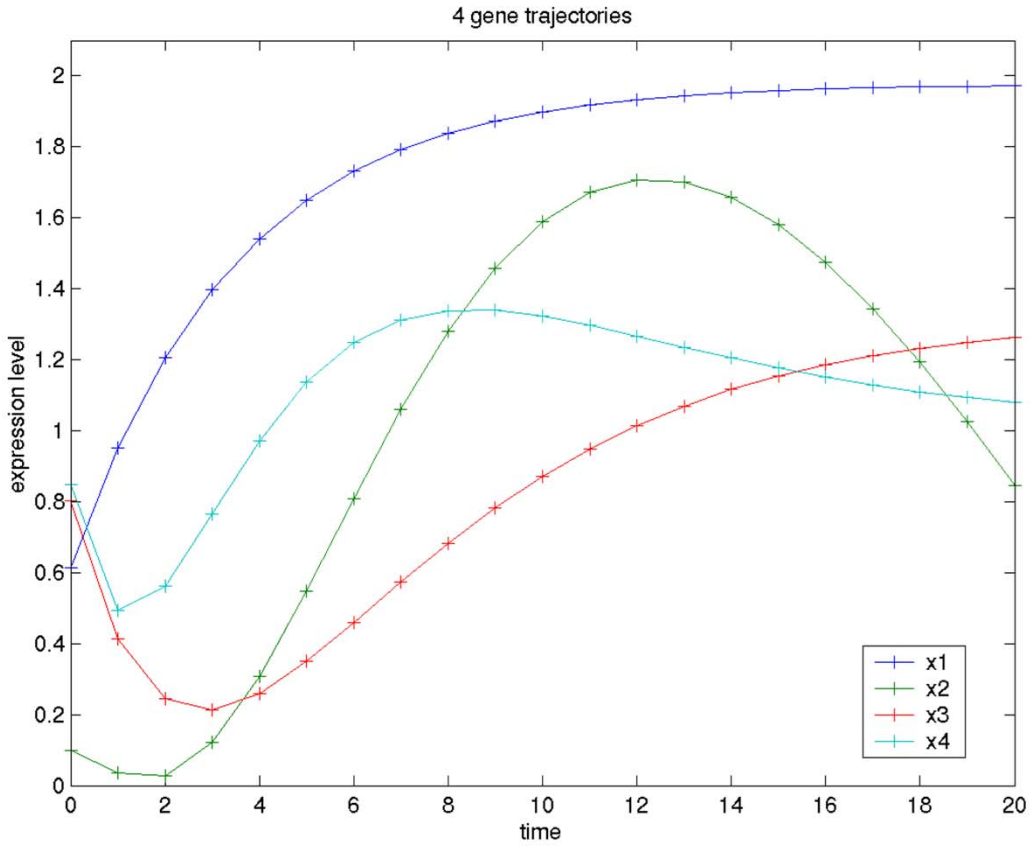
DNA Microarrays



- Experiment design
- Noise reduction
- Normalization
- …
- Data analysis

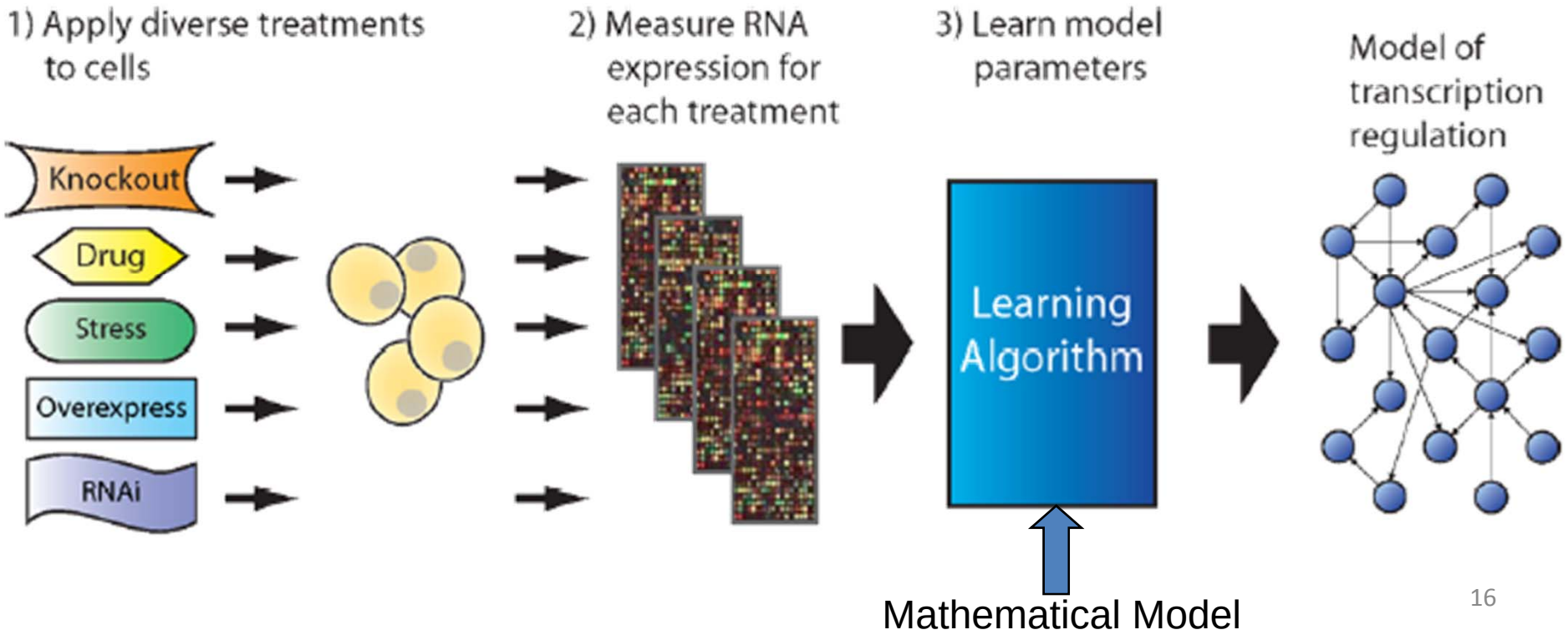
- Time series (e.g. cell cycle)
- Single time point (e.g. steady state)

Time Course Data



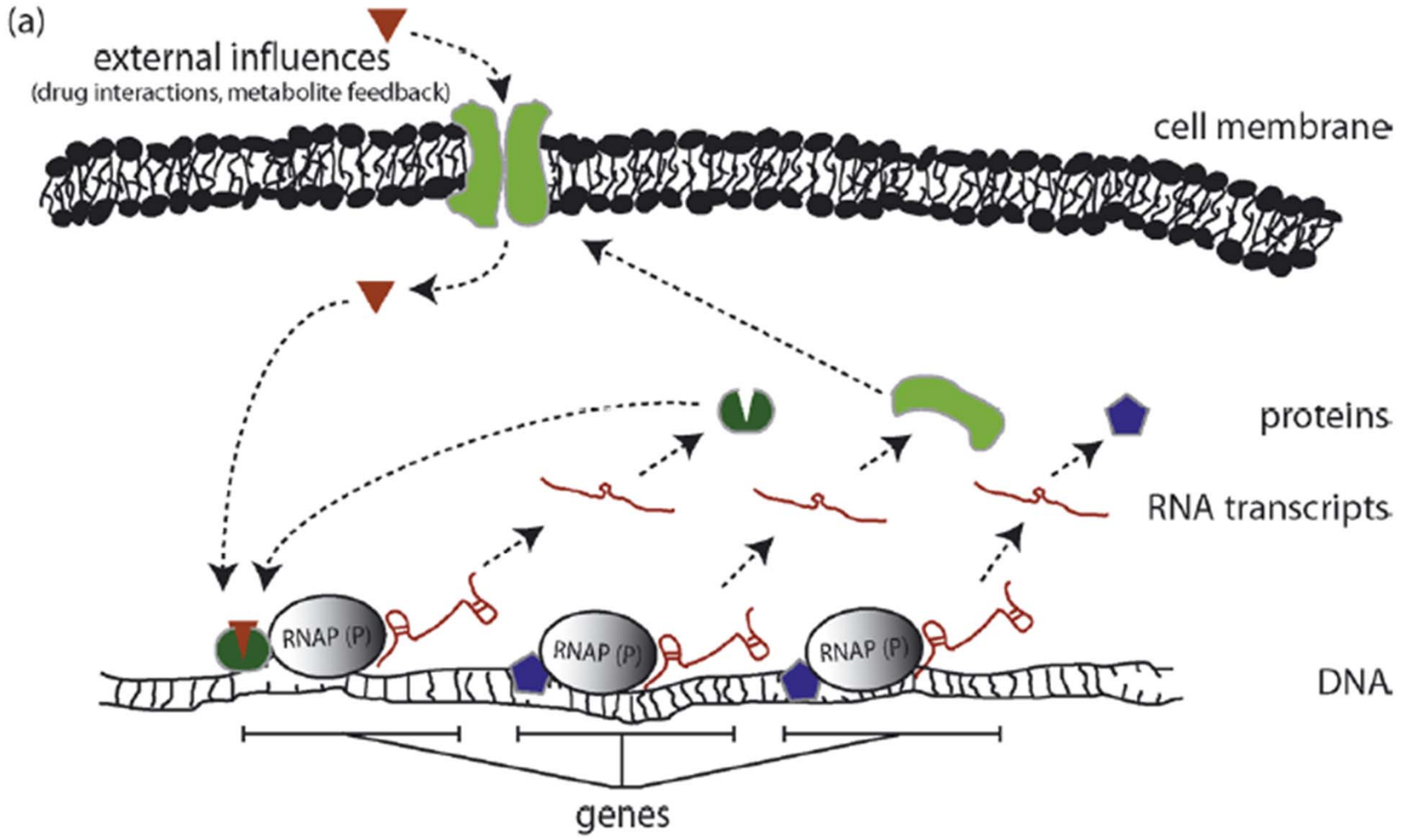
Basic idea

Goal: Infer structure and function of GRN from expression data



Outline

- Gene regulatory network modeling
  - Co-expression
  - Boolean networks
  - Bayesian models
  - Differential equations
- Gene regulatory network inference
  - GRNInfer
  - GNTInfer
  - GNMInfer
  - A detailed example

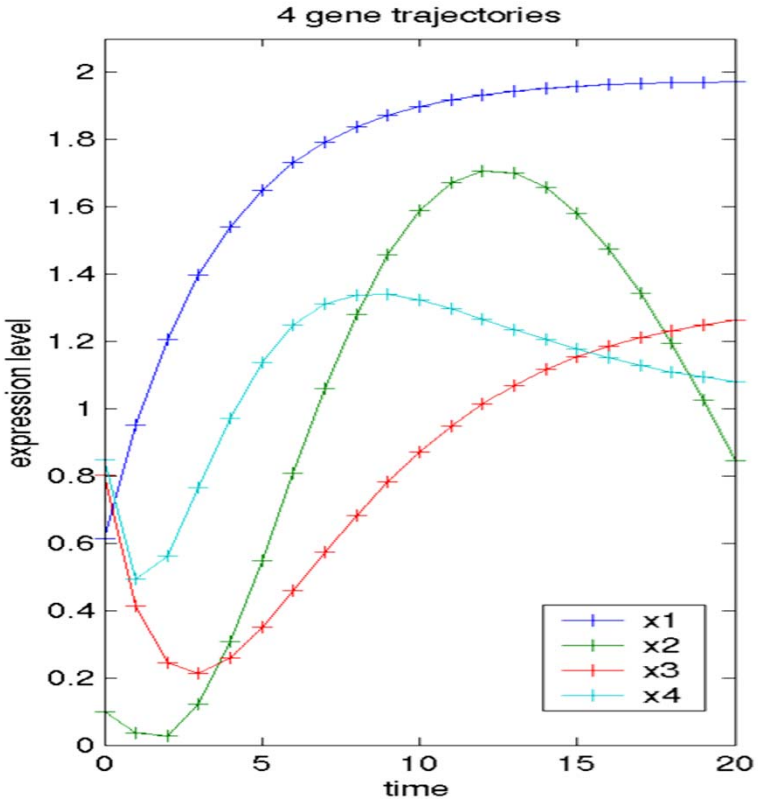


Only mRNA can be measured!

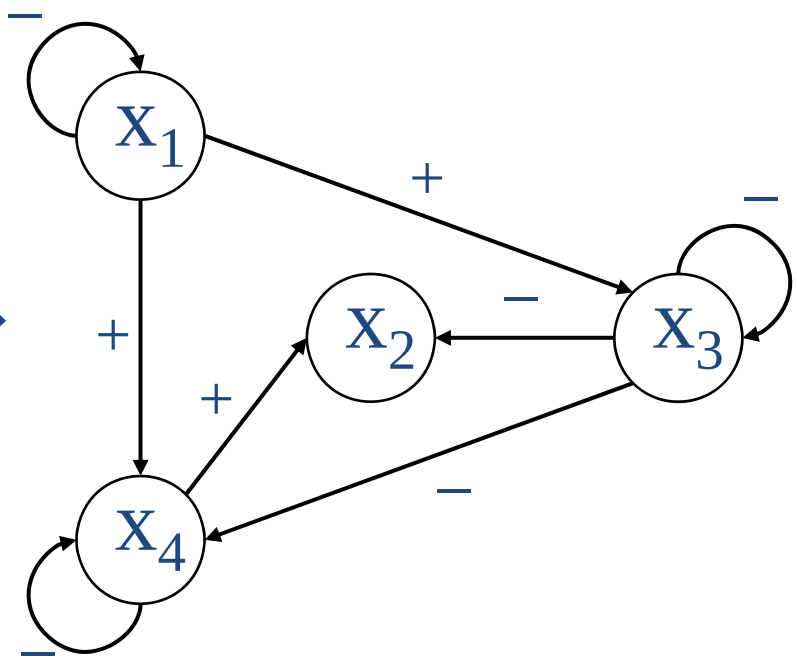
(b)

Gene regulatory network model

Model can not explicitly represents proteins and metabolites because only RNA can be measured



?



Time series

Gene network

Gene Expression Matrix

Given an experiment with *m* genes and *n* assays we produce a matrix *X* where:

x_{ij} = expression level of the i^{th} gene in the j^{th} assay.

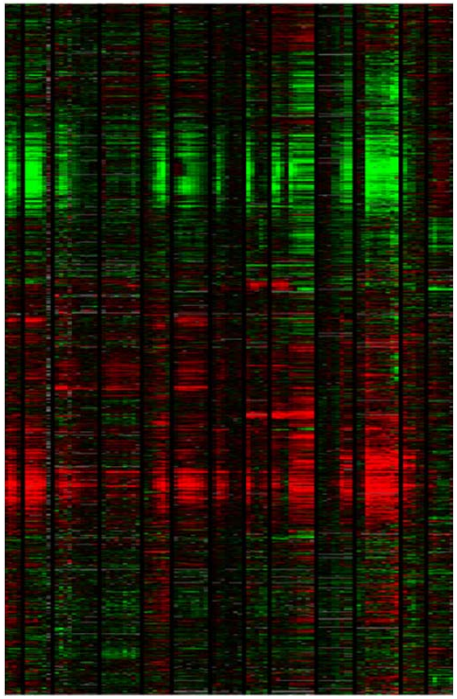
$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

g_i = Transcriptional response of the i^{th} gene

a_j = Expression profile of the j^{th} assay

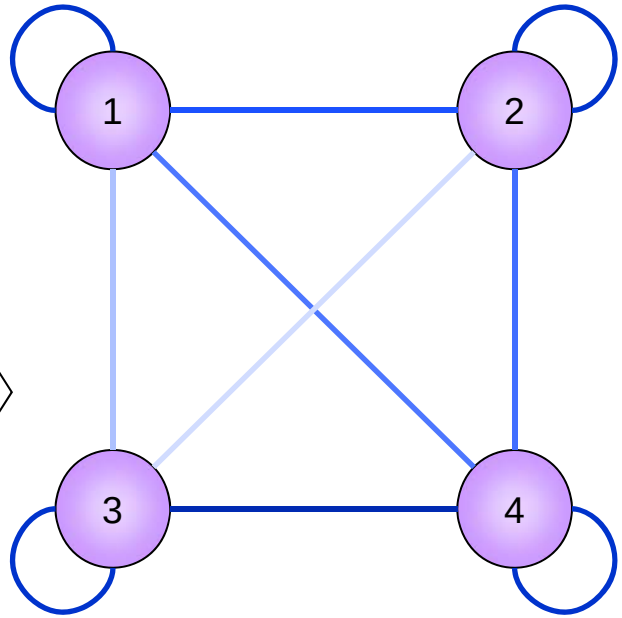
Correlation

- Gene expression



Gasch et al., 2000

$x_1 = (0.2, 2.4, 1.5, \ldots)$
$x_2 = (0.8, 2.2, 1.5, \ldots)$
$x_3 = (4.3, 0.1, 7.5, \ldots)$
…
$sim(x_1, x_2) = 0.62$
$sim(x_1, x_3) = -0.58$
…

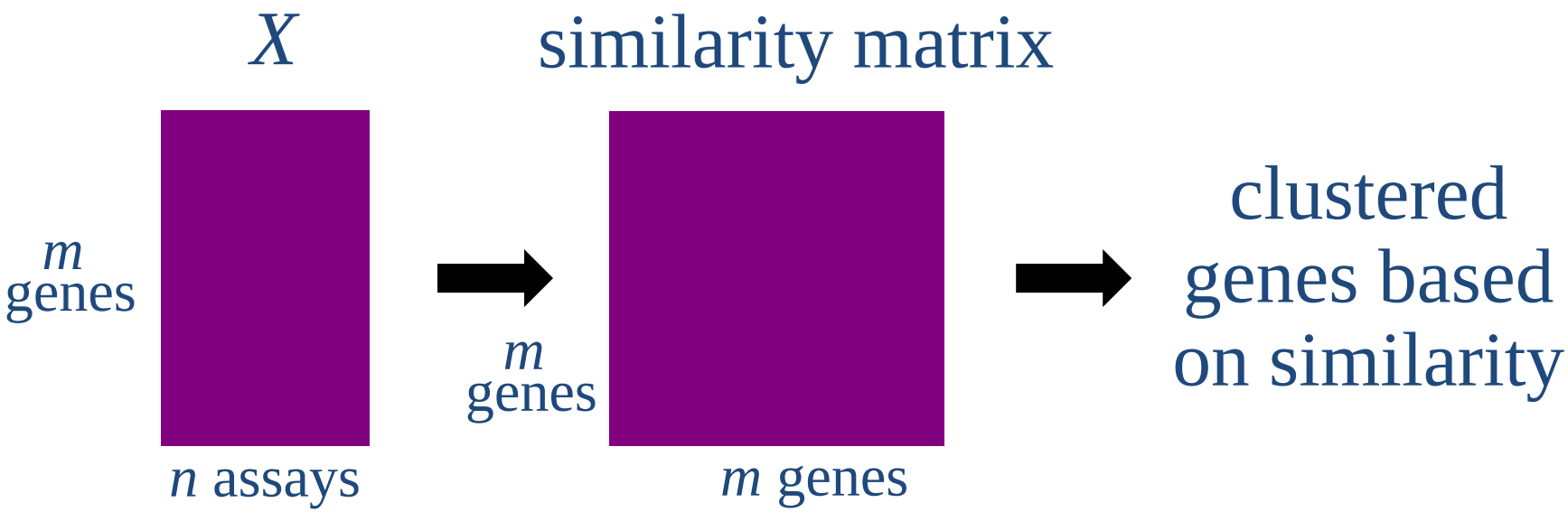


Similarity scale:
1 -1

Goals of Clustering

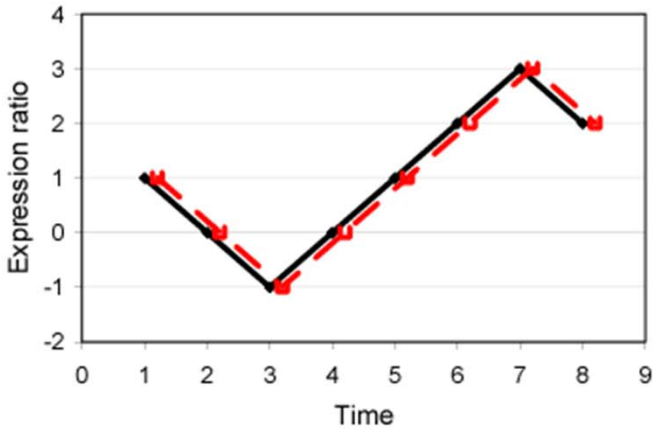
- Clustering genes:
  - Classify genes by their transcriptional response and get an idea of how groups of genes are regulated.
  - Potentially infer functions of unknown genes.
  - Construct relevance network (Co-regulation)
- Clustering assays:
  - Classify diseased versus normal samples by their expression profile.
  - Track the expression levels at different stages in the cell.
  - Study the impact of external stimuli.

Clustering Genes

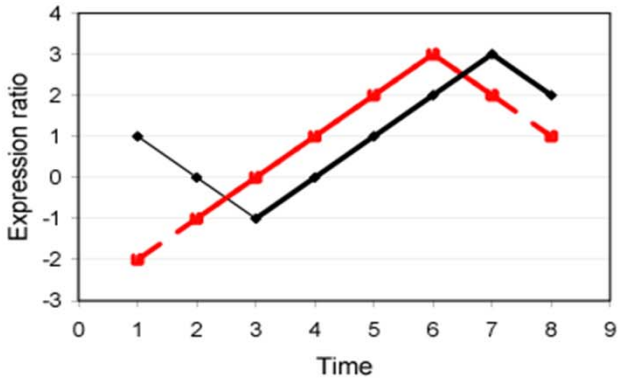


Simultaneous

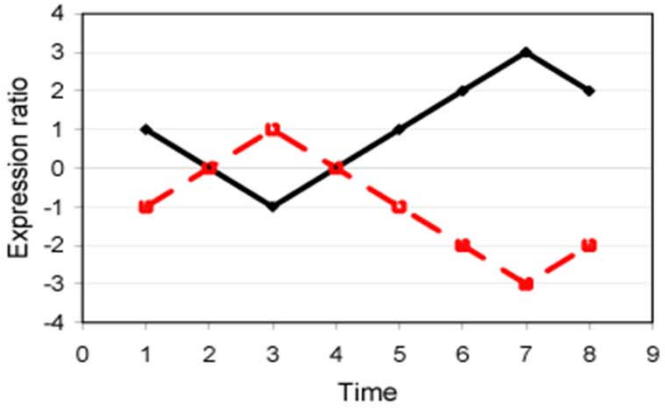
Traditional Global Correlation



Time-Shifted



Inverted

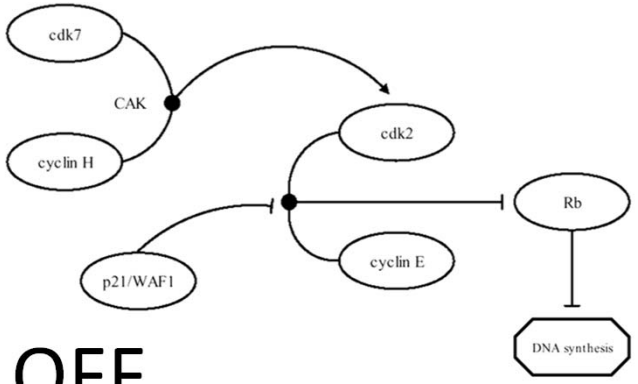


Local Clustering algorithm identifies further (reasonable) types of expression relationships

(Algorithm adapted from local sequence alignment)

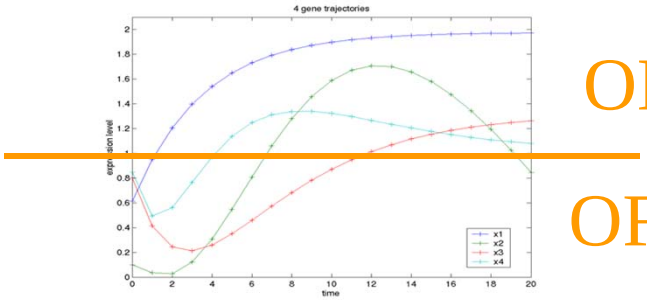
[Qian et al]

Boolean Networks



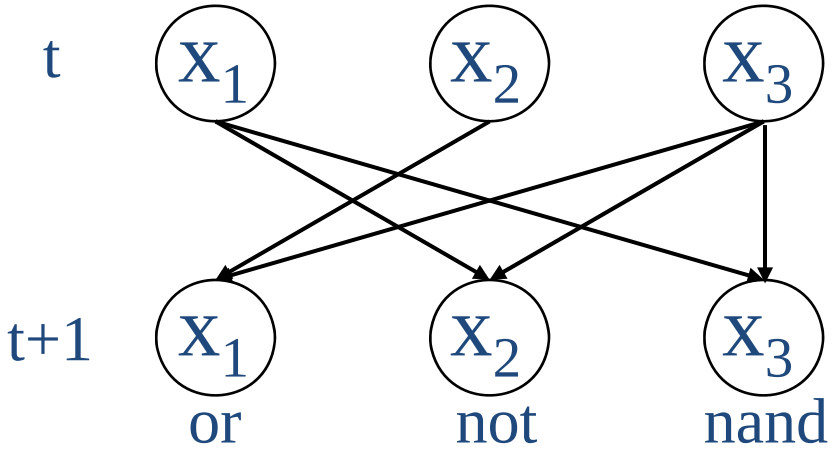
- Genes are assumed to be ON or OFF.
- At any given time, combining the gene states gives a *gene activity pattern* (GAP).
- Given a GAP at time *t*, a deterministic function (a set of logical rules) provides the GAP at time *t +1*.
- GAPs can be classified into *attractor* and *transient* states.

Boolean Network



ON

OFF



t	0	1	2	3	4
x_1	1	1	0	1	1
x_2	1	0	0	0	0
x_3	1	0	1	1	0

transient

attractors

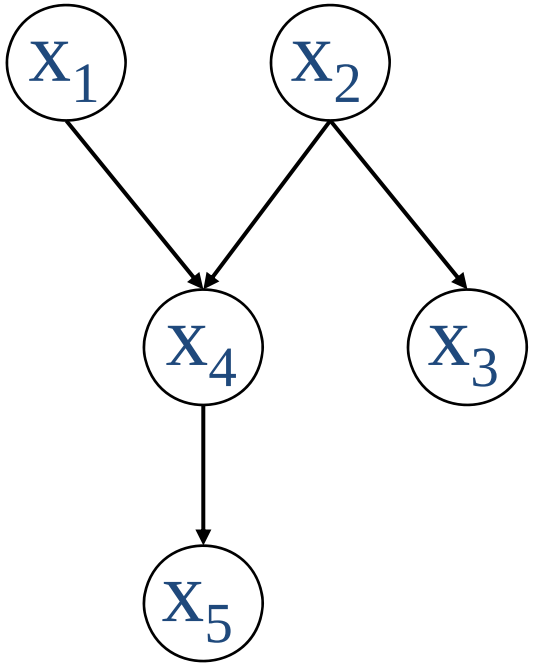
Issues with Boolean Networks

- Gene trajectories are continuous and modeling them as ON/OFF might be inadequate.
- A deterministic set of logical rules forces a very stringent model.
  - It doesn't allow for external input.
  - Very susceptible to noise.
- Probability Boolean Networks aims at fixing some of these issues by combining multiple sets of rules.

Bayesian Networks

- A gene regulatory network is represented by directed acyclic graph:
  - Vertices correspond to genes.
  - Edges correspond to direct influence or interaction.
- For each gene x_i, a conditional distribution $p(x_i \mid \text{ancestors}(x_i))$ is defined.
- The graph and the conditional distributions, uniquely specify the joint probability distribution.

Bayesian Network Example



Conditional distributions:

$p(x_1)$, $p(x_2)$, $p(x_3|\ x_2)$,
$p(x_4|\ x_1,x_2)$, $p(x_5|\ x_4)$

$$p(X) = p(x_1)\ p(x_2|\ x_1)\ p(x_3|\ x_1,x_2)\ p(x_4|\ x_1,x_2,\ x_3)\ p(x_5|\ x_1,x_2,\ x_3,x_4)$$

$$p(X) = p(x_1)\ p(x_2)\ p(x_3|\ x_2)\ p(x_4|\ x_1,x_2)\ p(x_5|\ x_4)$$

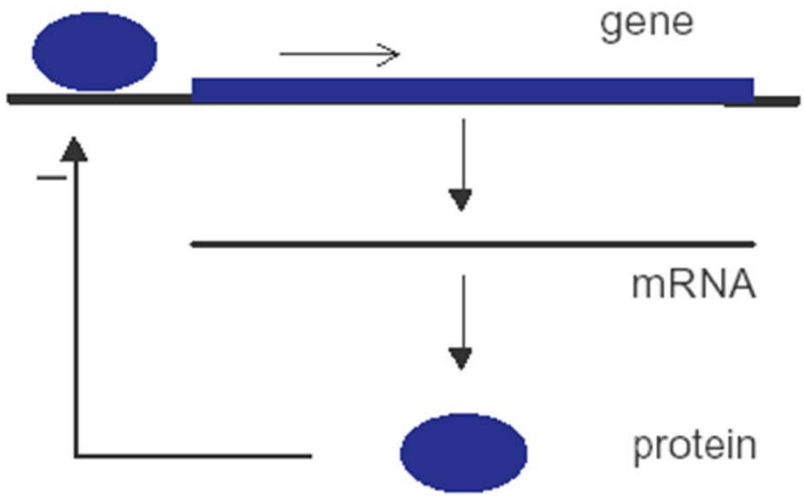
Learning Bayesian Models

- Using gene expression data, the goal is to find the Bayesian network that best matches the data.
- Recovering optimal conditional probability distributions when the graph is known is “easy”.
- Recovering the structure of the graph is NP-hard.

Issues with Bayesian Models

- Computationally intensive.
- Requires lots of data.
- Does not allow for feedback loops which play an important role (Network Motifs).
- Does not make use of the temporal aspect of the data.
- Dynamical Bayesian Networks aim at solving some of these issues but they require even more data.

Differential Equation Model



x_1 = mRNA concentration

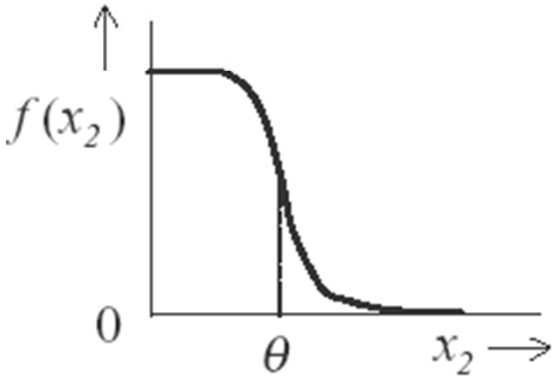
x_2 = protein concentration

$$\dot{x}_1 = \kappa_1 f(x_2) - \gamma_1 x_1$$

$$\dot{x}_2 = \kappa_2 x_1 - \gamma_2 x_2$$

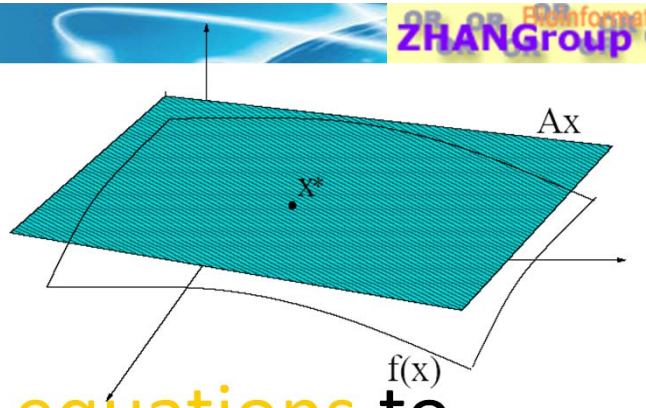
$\kappa_1, \kappa_2 > 0$, production rate constants

$\gamma_1, \gamma_2 > 0$, degradation rate constants



$$f(x_2) = \frac{\theta^n}{\theta^n + x_2^n}, \quad \theta > 0 \text{ threshold}$$

Linearization



- Typically uses linear differential equations to model the gene trajectories:

  $dx_i(t) / dt = a_0 + a_{i,1}\, x_1(t) + a_{i,2}\, x_2(t) + \ldots + a_{i,n}\, x_n(t) + u(t)$

- Reasons for that choice:
  - lower number of parameters implies that we are less likely to over fit the data
  - sufficient to model complex interactions between the genes

Issues with Differential Equations

- Even under the simplest linear model, there are *m(m+1)* unknown parameters to estimate:
  - *m(m-1) directional* effects
  - *m self* effects
  - *m constant effects*
- Number of data points is *m* and we typically have that *n* << *m* (few time-points).
- Extra constraints must be incorporated into the model such as:
  - Sparse structure of the network
  - Other prior information

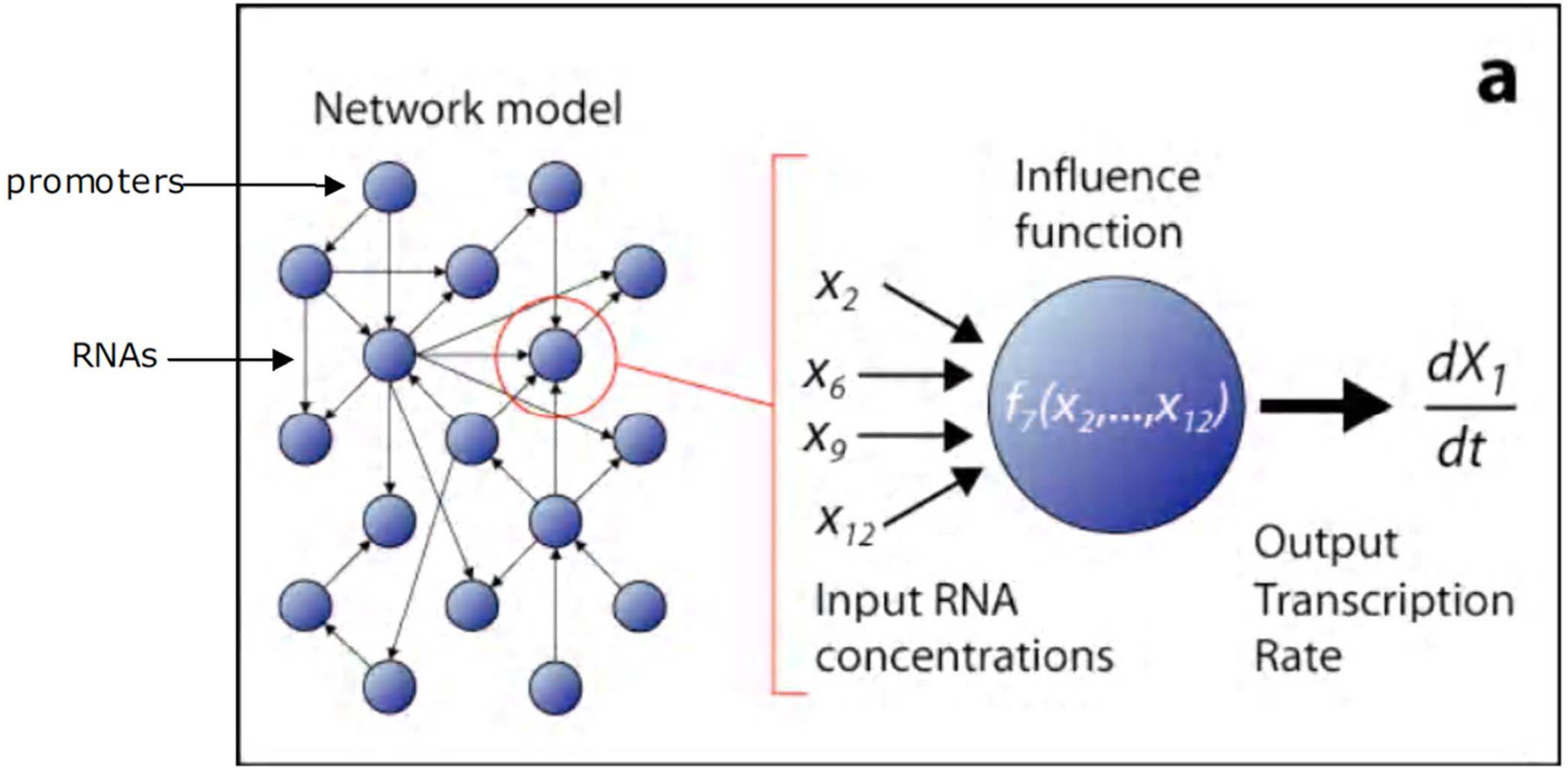
References

- Boolean Networks
  - Kauffman (1993), The Origins of Order
  - Lian et al. (1998), *PSB*, 3: 18-29.
- Bayesian Networks
  - Friedman et al. (2000), *RECOMB 2000*.
  - Hartemink et al. (2001), *PSB*, 6: 422-433.
- Differential Equations
  - Chen et al. (1999), *PSB*, 4: 29-40.
  - D'haeseleer et al. (1999), *PSB*, 4: 41-52.
  - Yeung et al. (2002), *PNAS*, 99(9): 6163-6168.
- Literature Review
  - De Jong (2002). *JCB*, 9(1): 67-103.
  - Gardner (2005). Physics of Life Reviews 2 :65–88.

Outline

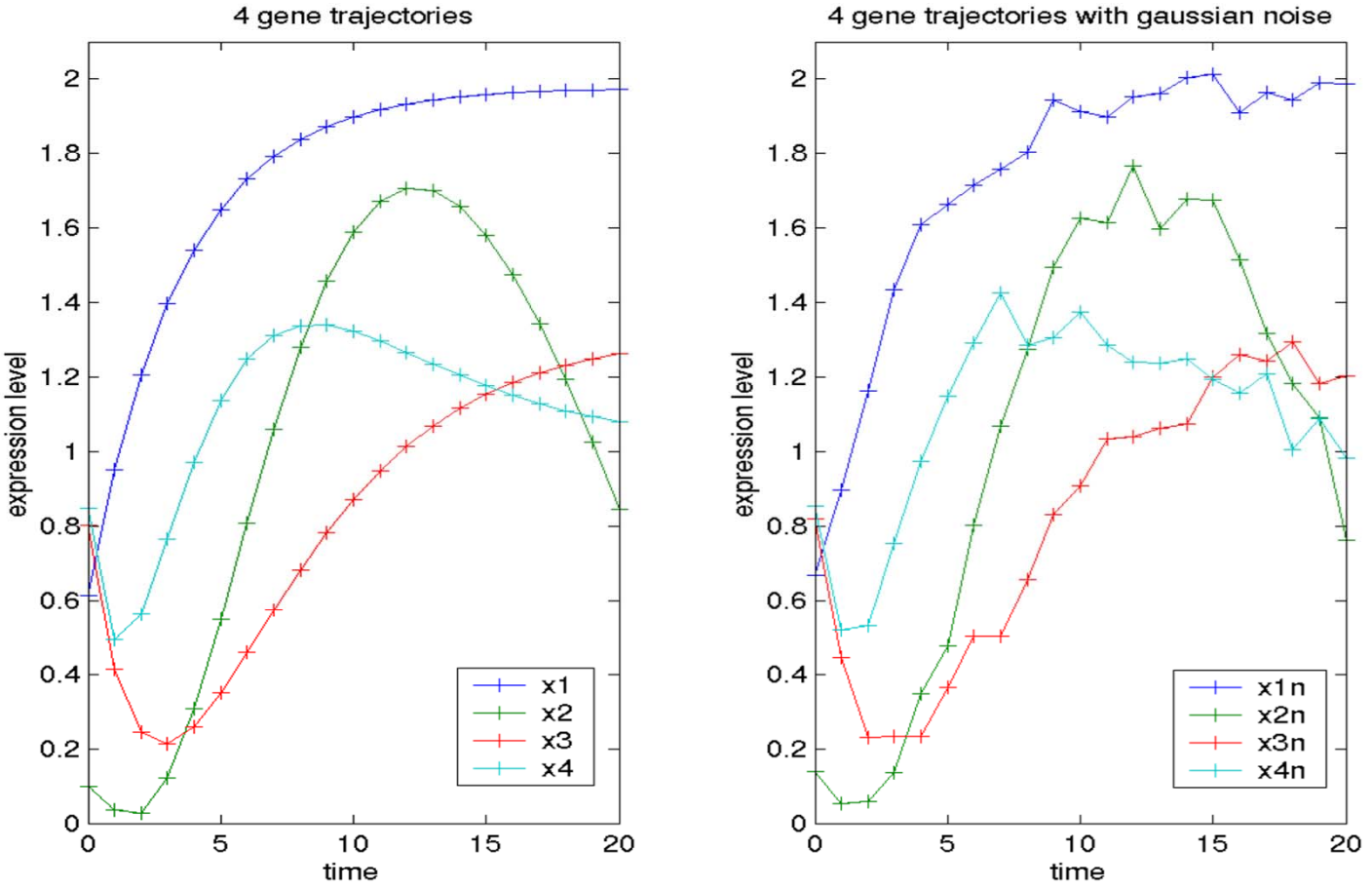
- Gene network modeling
  - Co-expression
  - Boolean networks
  - Bayesian models
  - Differential equations
- Gene regulatory network inference
  - GRNInfer
  - GNTInfer
  - GNMInfer
  - A detailed example

ODE model



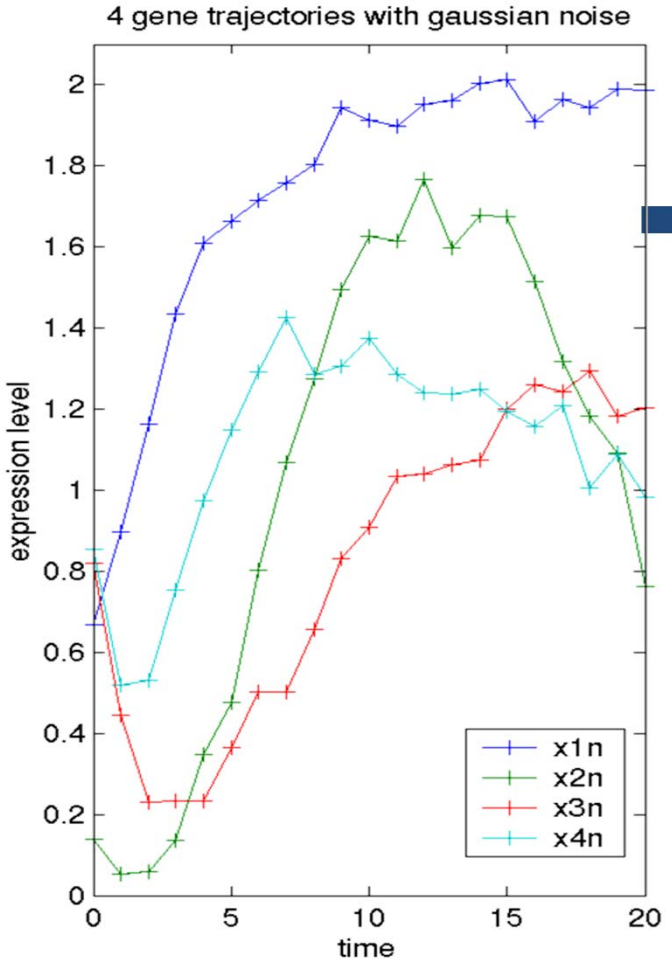
$$dX_1/dt = f_7(X_2,\ldots) = a_2 X_2 + a_6 X_6 + a_9 X_9 + a_{12} X_{12}$$

Noise

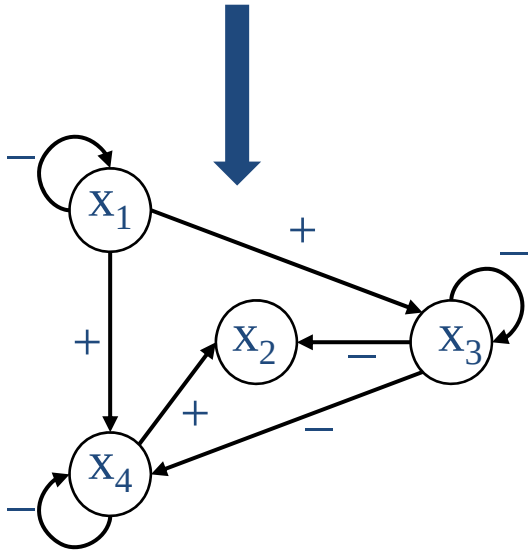


We add gaussian noise to model errors.

Network Inference



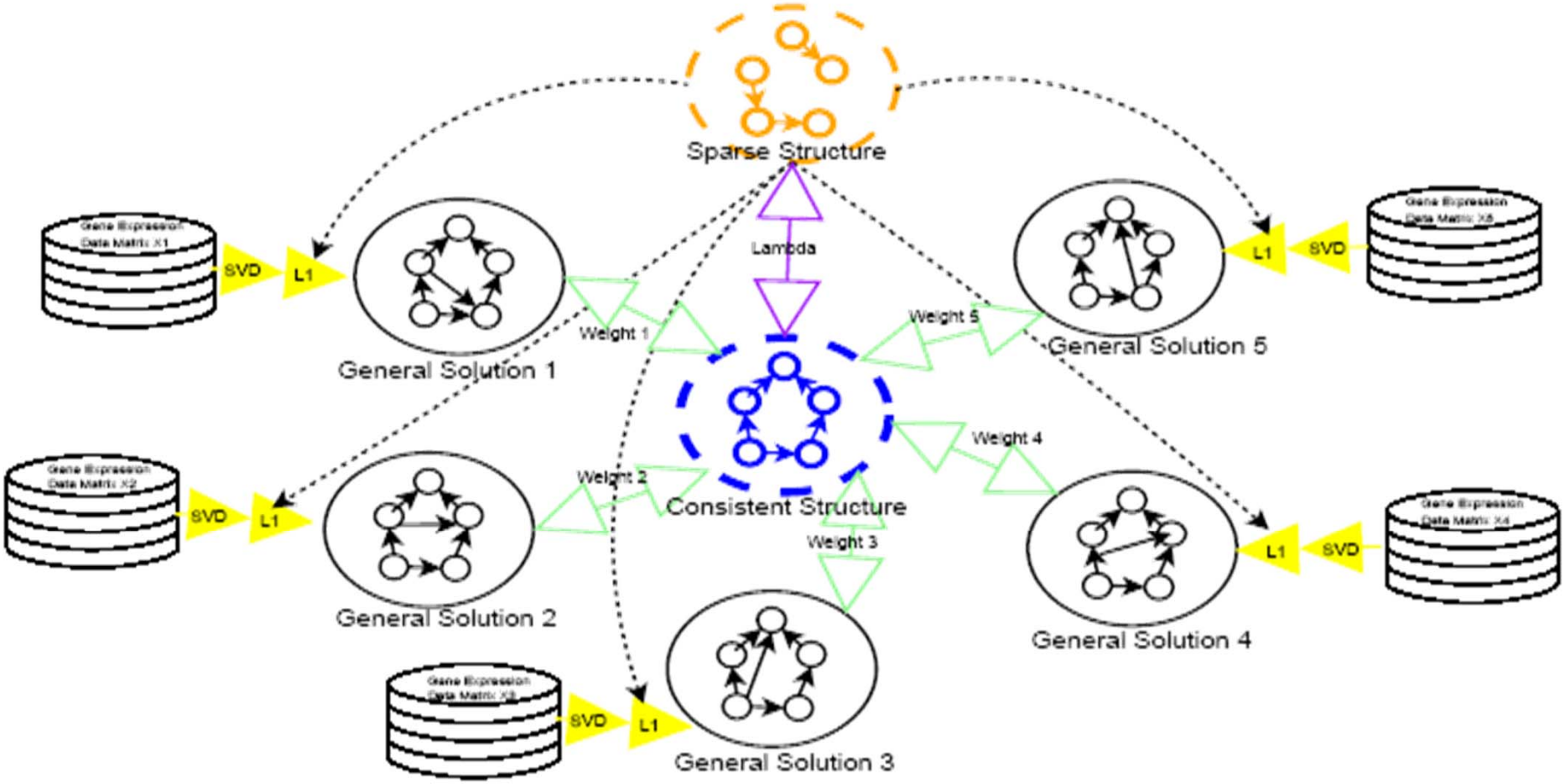
	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$	$a_{3,i}$	$a_{4,i}$
x_1	.431	-.248	0	0	0
x_2	0	0	0	-.473	.374
x_3	-.427	.376	0	-.241	0
x_4	0	.435	0	-.315	-.437



GRNInfer (Gene Regulatory Network reconstruction tool)

- A single dataset consists of relatively few time points (less than 20) but a large number of genes (in thousands)
- Multiple Gene expression datasets are generated by different groups worldwide are increasingly accumulated on many species
- Combining and further exploiting multiple datasets in an integrative and systematic manner, the scarcity of data can be greatly alleviated.
- A more accurate reconstruction of GN can be expected.
- Simply arranging multiple time-course datasets into a single time-course dataset is Inappropriate for GN inference due to data normalization issues and lack of temporal relationships among datasets.
- A biological gene network is expected to be sparse

GRNInfer scheme



General solution of a single dataset

$$\dot{x}(t) = f(x(t))$$

$$\dot{x}(t) = Jx(t) + b(t), \quad t = t_1, \ldots, t_m$$

$$\dot{X} = JX + B$$

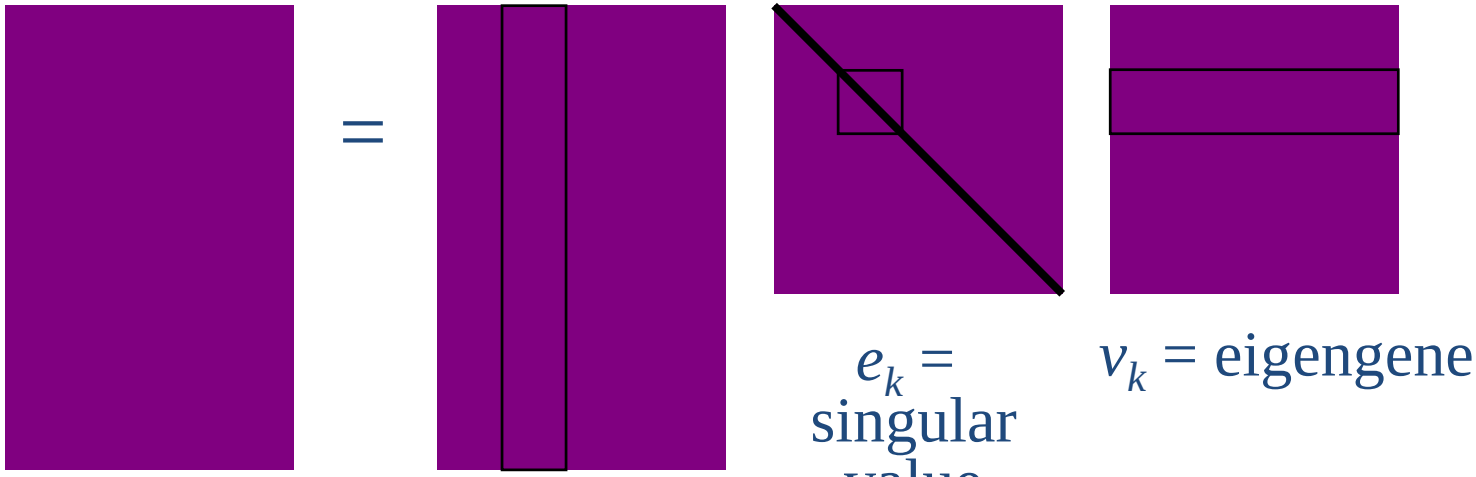
Infer a linear model

- Curse of dimension: #of experiments <<#of variables m(20)<<n(6000)
- ➔Inference problem is undetermined
- How to recover J? (Infinitely many possible solutions➔ many network architecture fit the data)
- Find one possible solution as a particular solution (SVD Singular Value Decomposition)

$$J_{n\times n}X_{n\times m} = \dot{X}_{n\times m} - B_{n\times m}$$

Singular Value Decomposition

$$X^T_{mx\,n} = U_{m\,x\,n}\; E_{n\,x\,n}\; V^{\,T}_{\;n\,x\,n} \quad (m<<n)$$



gene expression matrix

u_k = eigenassay

e_k = singular value

v_k = eigengene

$$\hat{J} = (\dot{X} - B)UE^{-1}V^T$$

- SVD solution is the particular solution in the least square meaning

$$\hat{J} = argmin \| JX + B - \dot{X} \|_2$$

- General solution: affine space

$$J = \hat{J} + YV^T \qquad Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1l} & 0.0 & \cdots & 0.0 \\ y_{21} & y_{22} & \cdots & y_{2l} & 0.0 & \cdots & 0.0 \\ \cdots & & \cdots & & & & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nl} & 0.0 & \cdots & 0.0 \end{bmatrix}$$

- Y denotes all degrees of the freedom can be used to optimize some extra criterion
- For example the sparsity of J ➔Maximize the number of zeros in J
- Impose J=0 ➔ i. e.

$$\hat{J} = -YV^T$$

The general solution represents all of the possible networks that are consistent with the single microarray dataset, depending on arbitrary Y.

We will find the most consistent network structure $J = (J_{ij})_{n \times n}$ for all $k = 1, ..., N$, with consideration of sparse structure

Optimization model

$$\min_{Y,J} \quad \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{n} [\omega^k |J_{ij} - J_{ij}^k| + \lambda |J_{ij}|]$$

Decomposition Algorithm

STEP-0: Initialization. Obtain all of the particular solution $\hat{J}^k$ by SVD, and ω^k. Set initial value $J_{ij}(0) = 0$, $Y^k_{ij}(0) = 0$ and $J^k_{ij}(0) = \hat{J}^k$, and positive λ, ϵ. Set $q = 1$.

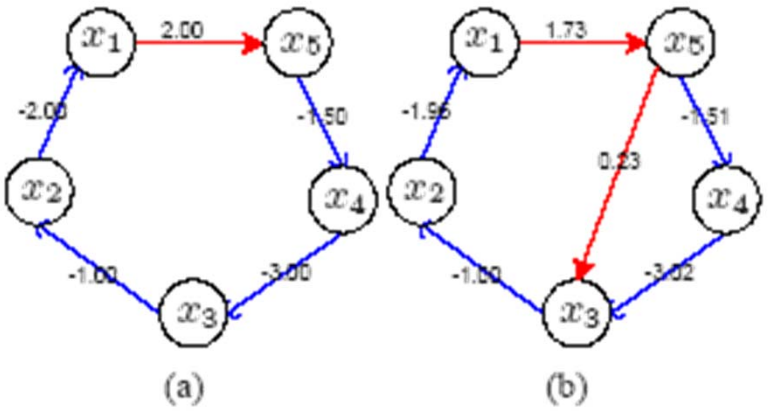
STEP-1: Set $J^k(q) = J^k(q-1) + Y^k(q)V_k^T$ and solve $y^k_{ij}(q)$ at iteration q by LP with $J(q-1)$ fixed, i.e. solve $Y^k(q) = (y^k_{ij}(q))_{m\times m}$ of the following subproblem for $k = 1, ..., N$ with $J(q-1)$ given ($y^k_{ij}(q) = 0$ if $j > l_k$)

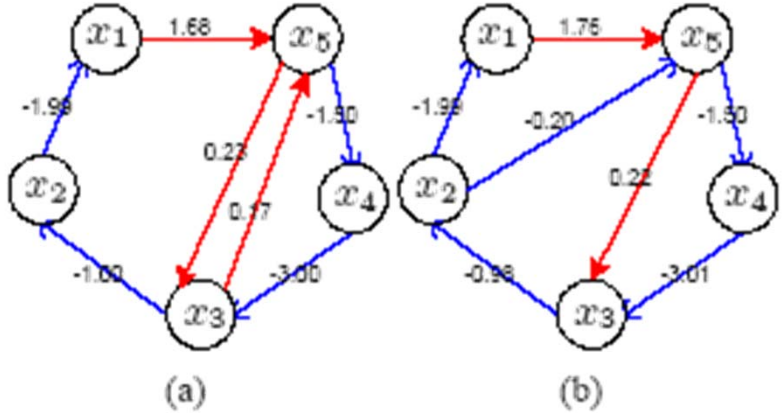
$$\min_{Y^k(q)} \sum_{i=1}^{n} \sum_{j=1}^{n} |J_{ij}(q-1) - J^k_{ij}(q)|$$

STEP-2: Solving $J_{ij}(q)$ at iteration q by LP with all of $y^k_{ij}(q)$ given, i.e. solve $J(q)$ of the following problem with all of $J^k(q)$ fixed.

$$\min_{J(q)} \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{n} [\omega^k |J_{ij}(q) - J^k_{ij}(q)| + \lambda |J_{ij}(q)|]$$

Simulated examples



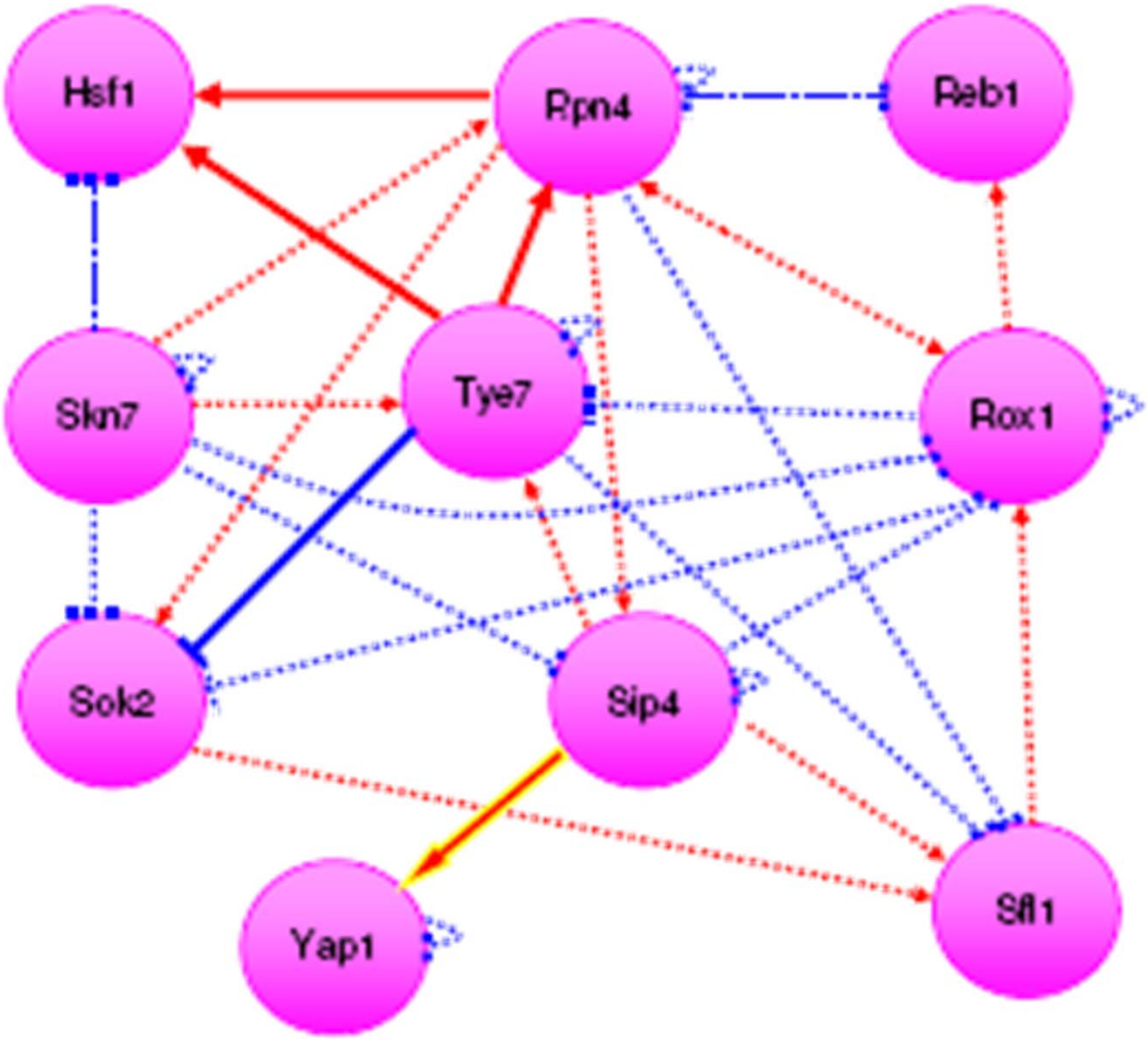


$$
\begin{aligned}
\dot{x}_1(t) &= -2x_2(t) &&+ \xi_1(t),\\
\dot{x}_2(t) &= -x_3(t) &&+ \xi_2(t),\\
\dot{x}_3(t) &= -3x_4(t) &&+ \xi_3(t),\\
\dot{x}_4(t) &= -1.5x_5(t) &&+ \xi_4(t),\\
\dot{x}_5(t) &= 2x_1(t) &&+ \xi_5(t),
\end{aligned}
$$

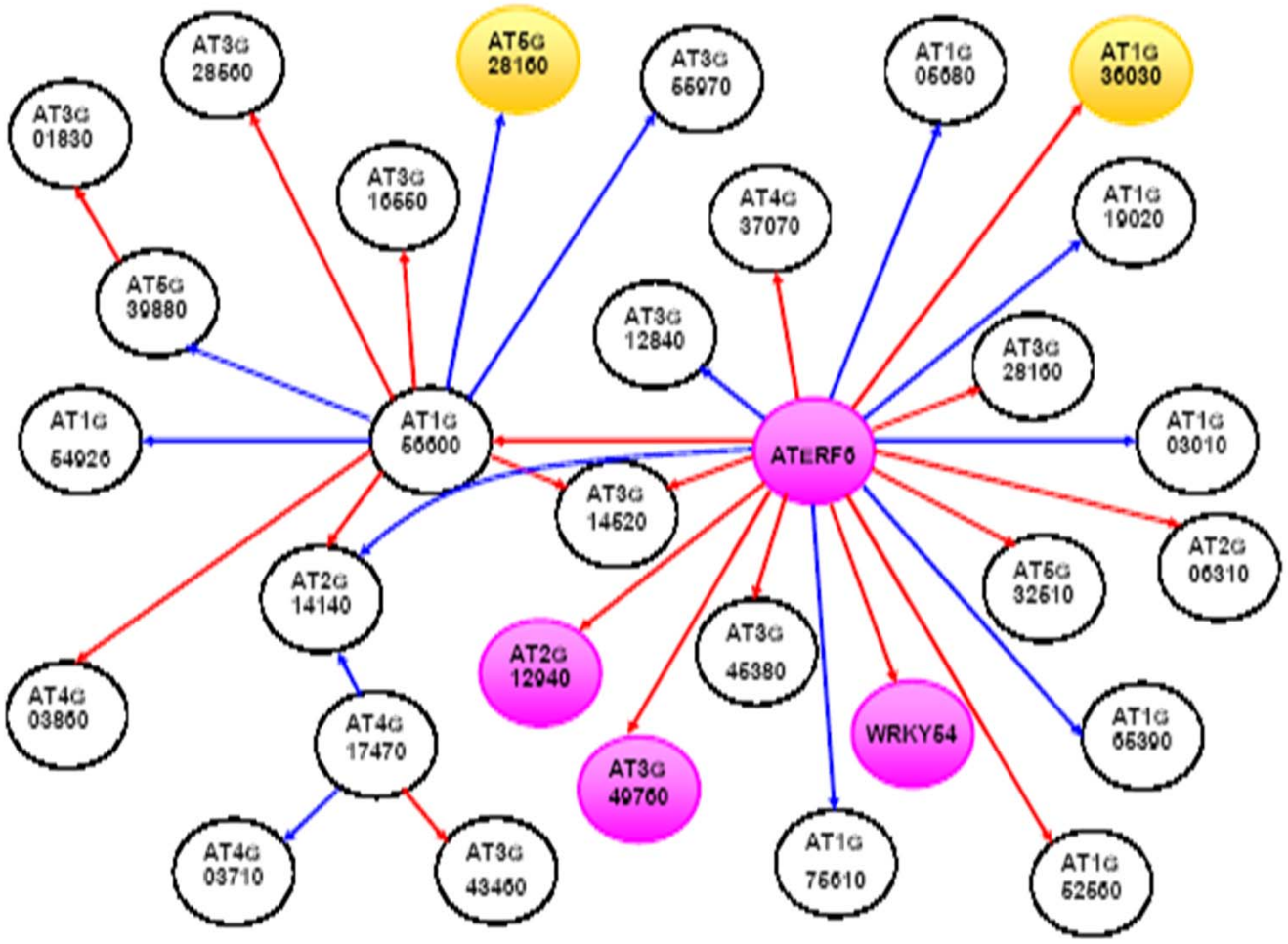
Without Noise

With Noise

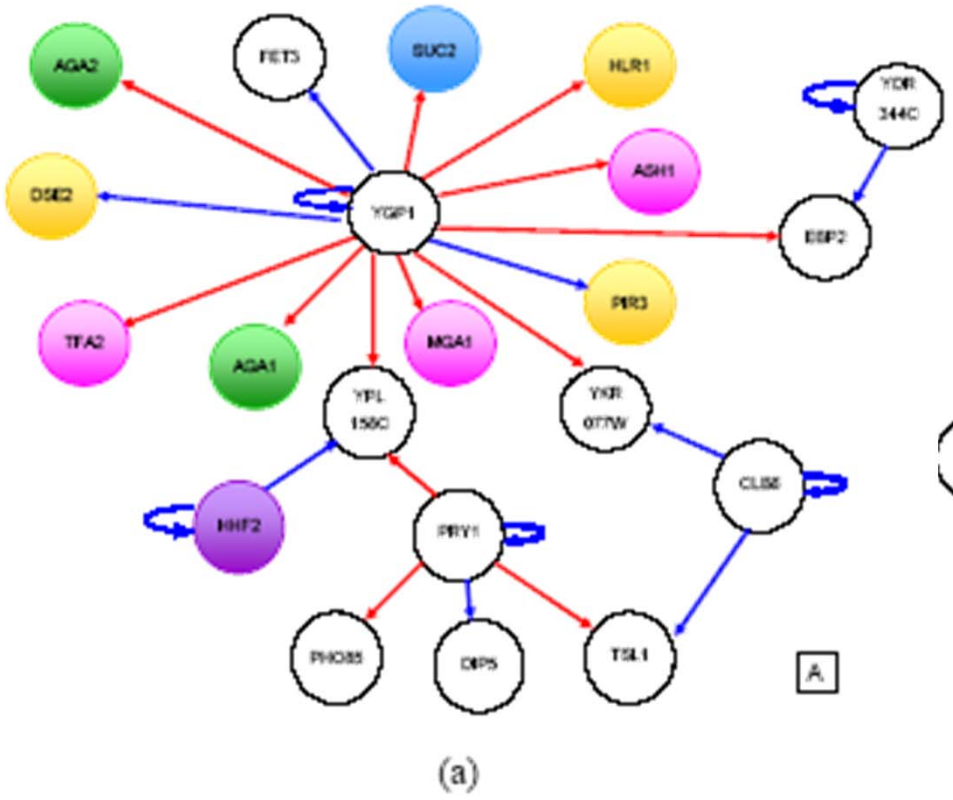
Heat-Shock Response for Yeast



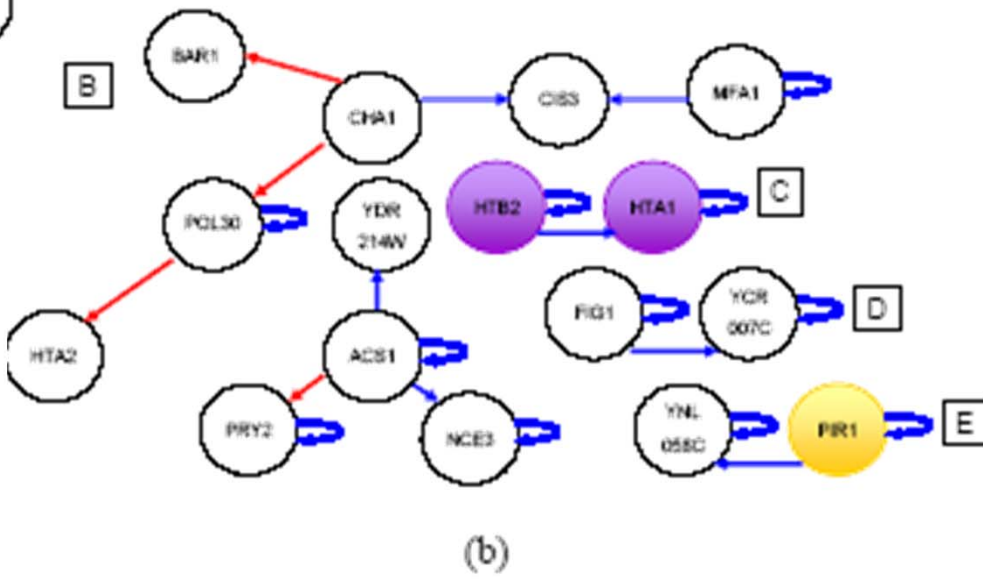
Yeast Cell Cycle (4 datasets)



Arabidopsis (9 Datasets)

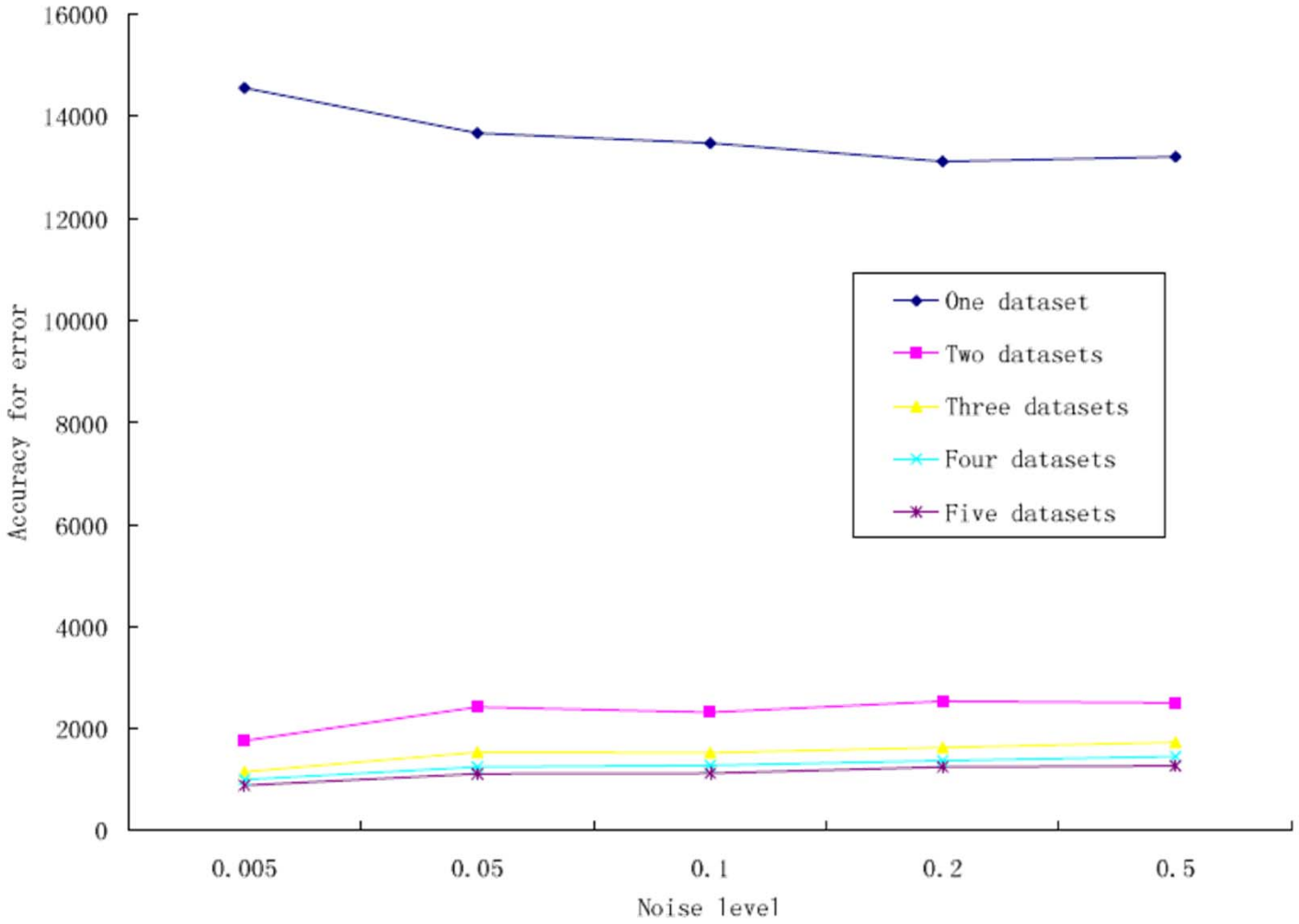


(a)

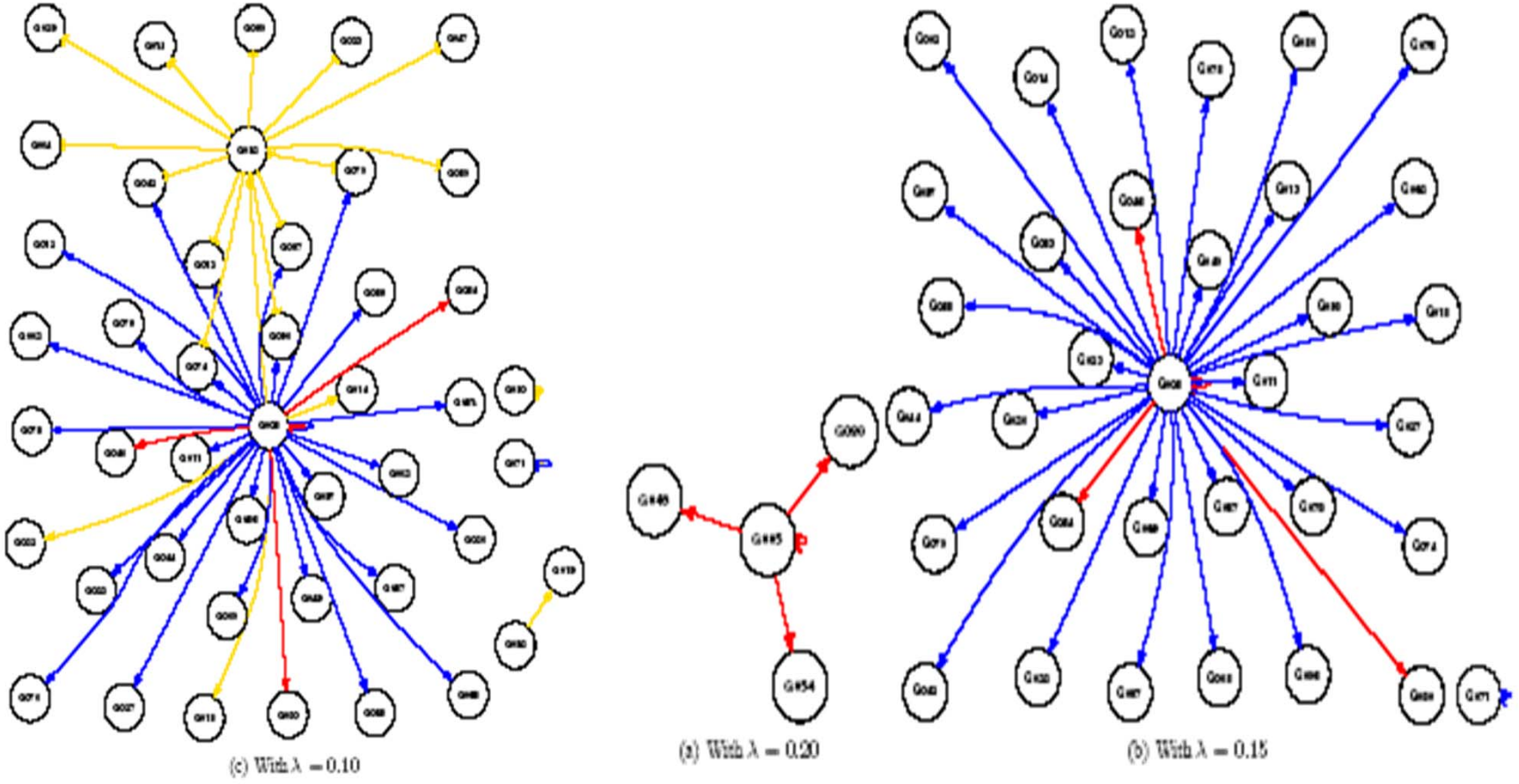


(b)

The advantage of multiple datasets



Consistent structure



(c) With $\lambda = 0.10$

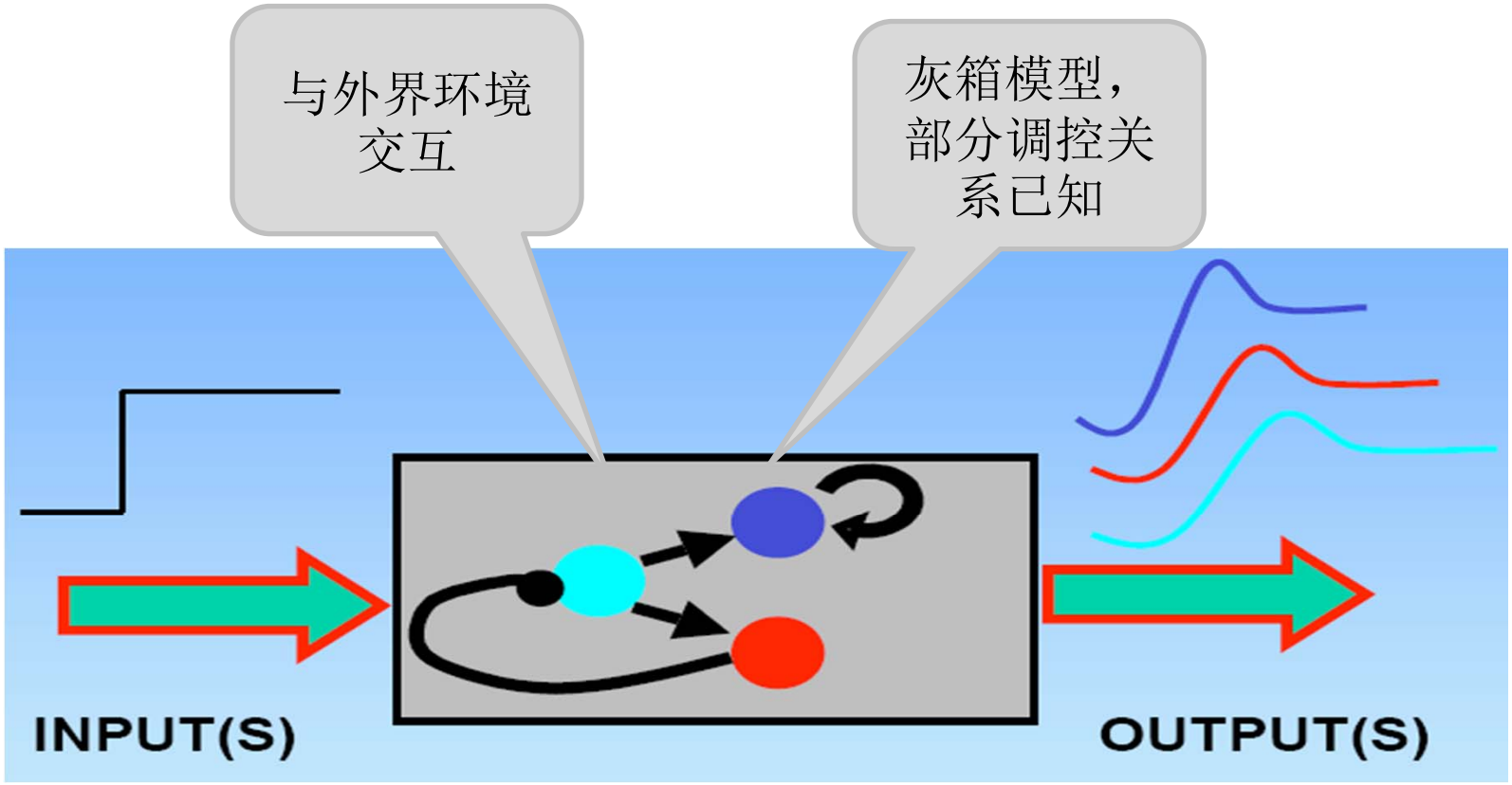
(a) With $\lambda = 0.20$

(b) With $\lambda = 0.15$

GNTInfer(Gene Network reconstruction tool with compound Targets)

- Include other available information derived from expression profile and from published literature so as to recover gene regulations in a more robust and reliable manner.
- Incorporate external inputs or perturbations into the formulation so that molecular targets (genes) can be identified in a systematic way.

考虑外部环境影响

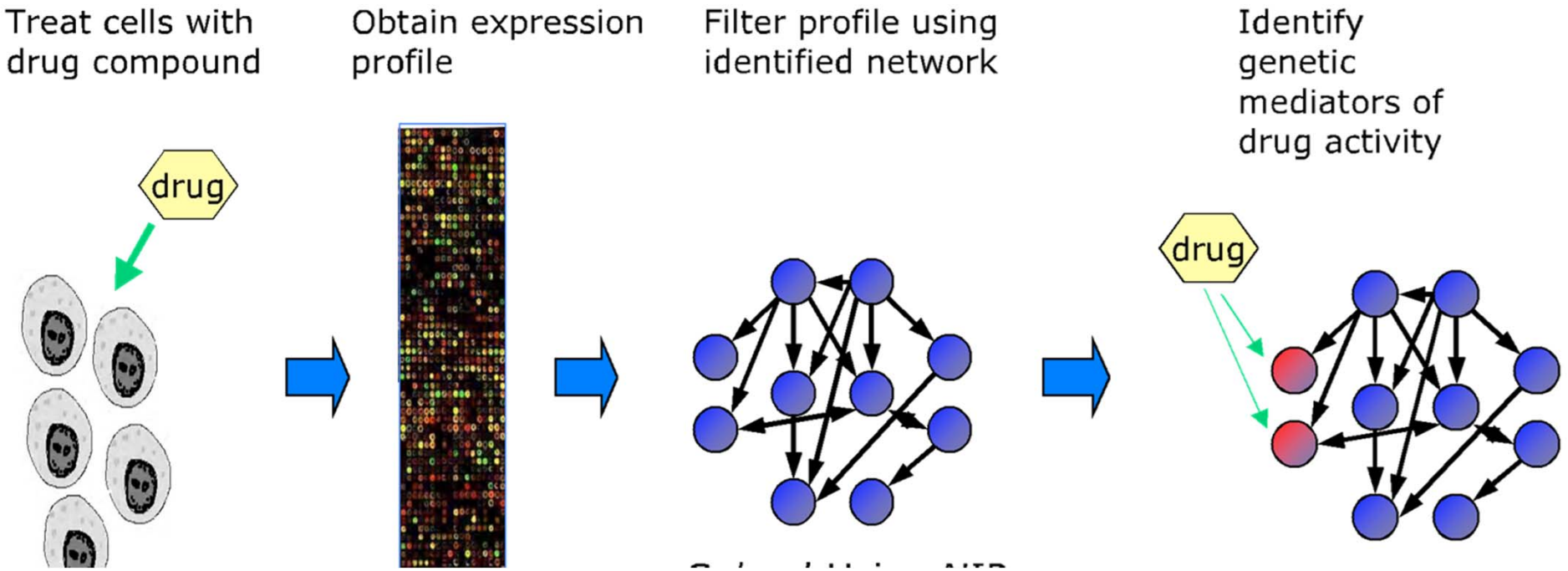


预测化合物的靶基因(compound Targets)

- 考虑外部输入或者扰动对基因调控网络的影响，用系统的方式识别他们的靶点基因

- 可以考虑的外部因素：
- 环境因素：温度、压力
- 药物或化合物
- 非编码RNA
- 基因敲除
- 其它

Identify compound Targets



数学表达

- 引入控制项

$$\dot{X} = J_{n\times n} X + P_{n\times s} C + \varepsilon$$

$$X(1), \cdots, X(m), C(1), , \cdots, C(t) \Rightarrow J_{n\times n}, P_{n\times s}$$

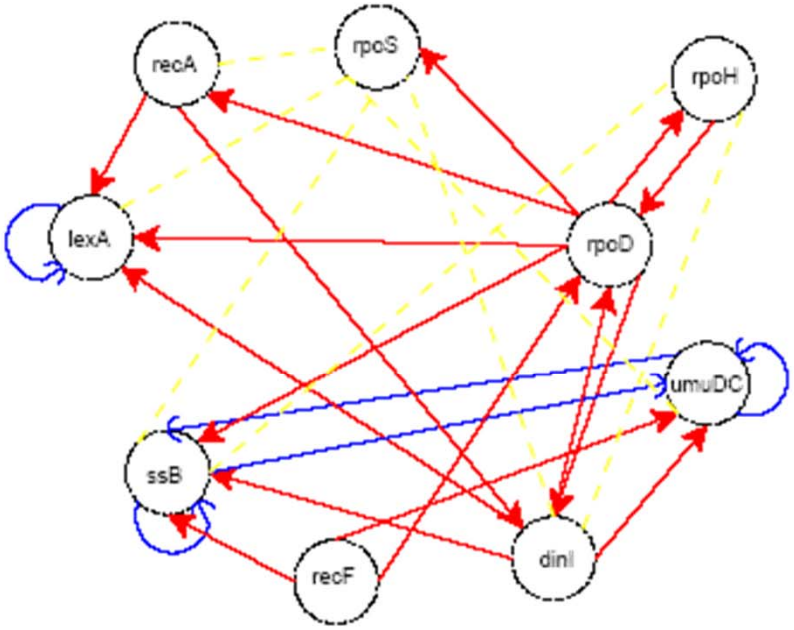
$$X(t) \in \mathrm{R}^{n}, C(t) \in \mathrm{R}^{s} \qquad m << n$$

- **P** 代表 **s** 个外部扰动对各个基因的影响
- 已知**X, C,** 求矩阵

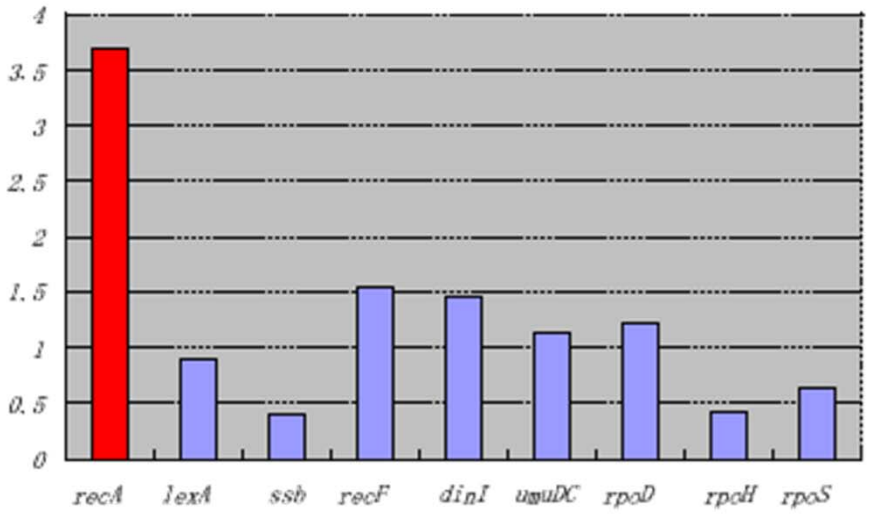
$$\dot{X} = [J, P] \begin{bmatrix} X \\ C \end{bmatrix}$$

E. Coli SOS Pathway

	recA	lexA	ssb	recF	dinI	umuDC	rpoD	rpoH	rpoS	Perturbation
recA	-0.0682	0.1149	0.0599	-0.0095	-0.0431	0.0000	**0.0173**	-0.0104	*0.0000*	0.1739
lexA	**0.0009**	**-0.1098**	0.0232	-0.0197	**0.0061**	0.0000	**0.0082**	0.0384	*0.0000*	0.0418
ssb	-0.0181	0.0188	**-0.0141**	**0.0279**	**0.0020**	**-0.0192**	**0.0018**	*0.0000*	*0.0000*	0.0187
recF	-0.0424	0.0015	0.0539	-0.0863	*0.0000*	-0.0090	-0.0005	0.0398	0.0000	0.0731
dinI	**0.0268**	0.0239	0.0538	0.0000	-0.0827	0.0769	**0.0177**	*0.0000*	*0.0000*	0.0689
umuDC	0.0000	0.0000	**-0.0527**	**0.0247**	**0.0280**	**-0.0705**	0.0000	0.0083	*0.0000*	0.0531
rpoD	-0.0525	0.0237	0.0145	**0.0009**	**0.0059**	0.0000	-0.0211	**0.0336**	*0.0000*	0.0578
rpoH	-0.0256	-0.0143	*0.0000*	-0.0111	*0.0000*	0.0335	**0.0127**	-0.0032	*0.0000*	0.0195
rpoS	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	**0.0101**	0.0091	-0.0274	0.0304



(a) Predicted network structure



(b) Predicted perturbation

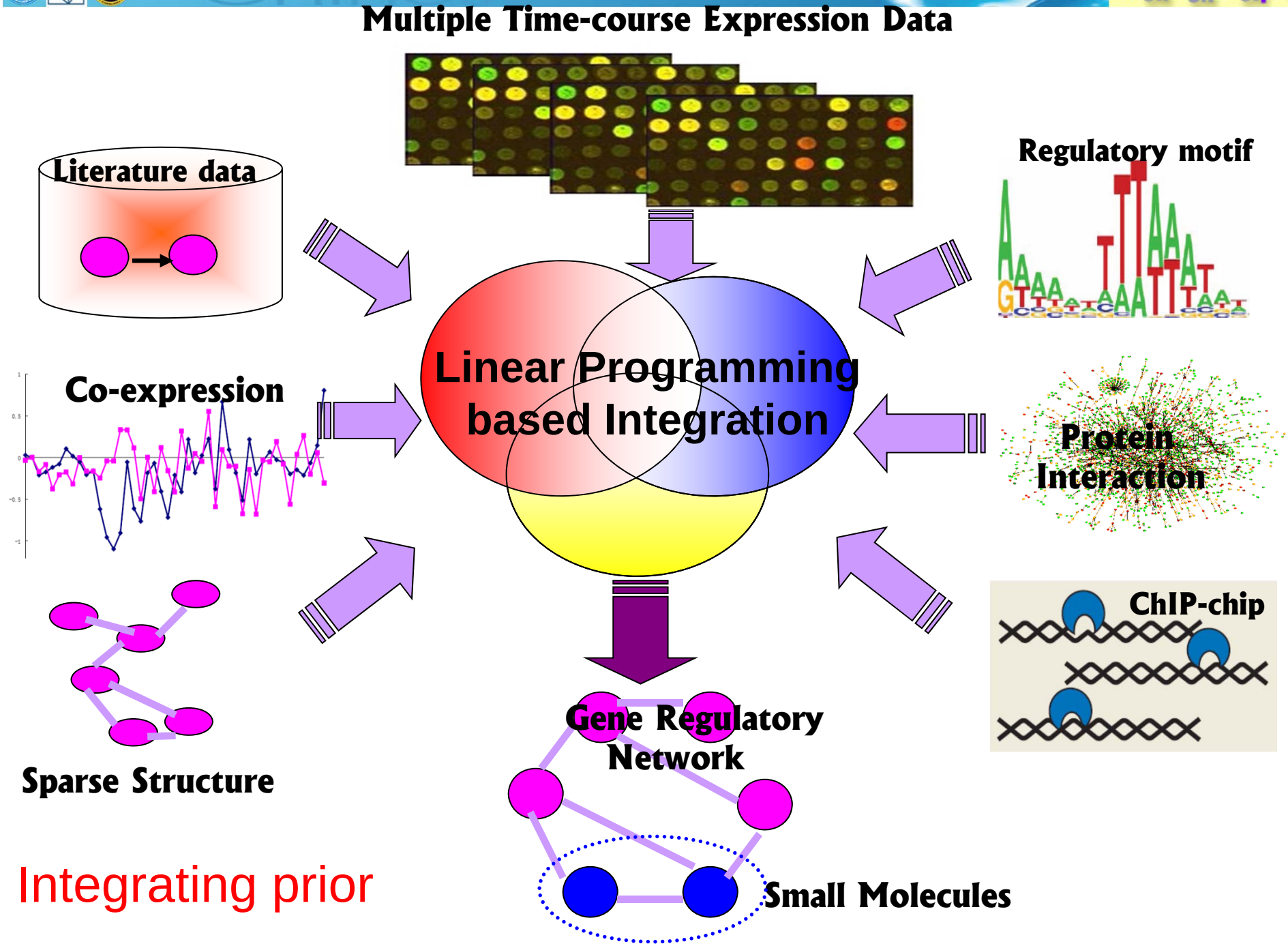
集成先验信息

1. 基因调控网络中的维度问题

2. 大量的关于基因调控网络的异源数据

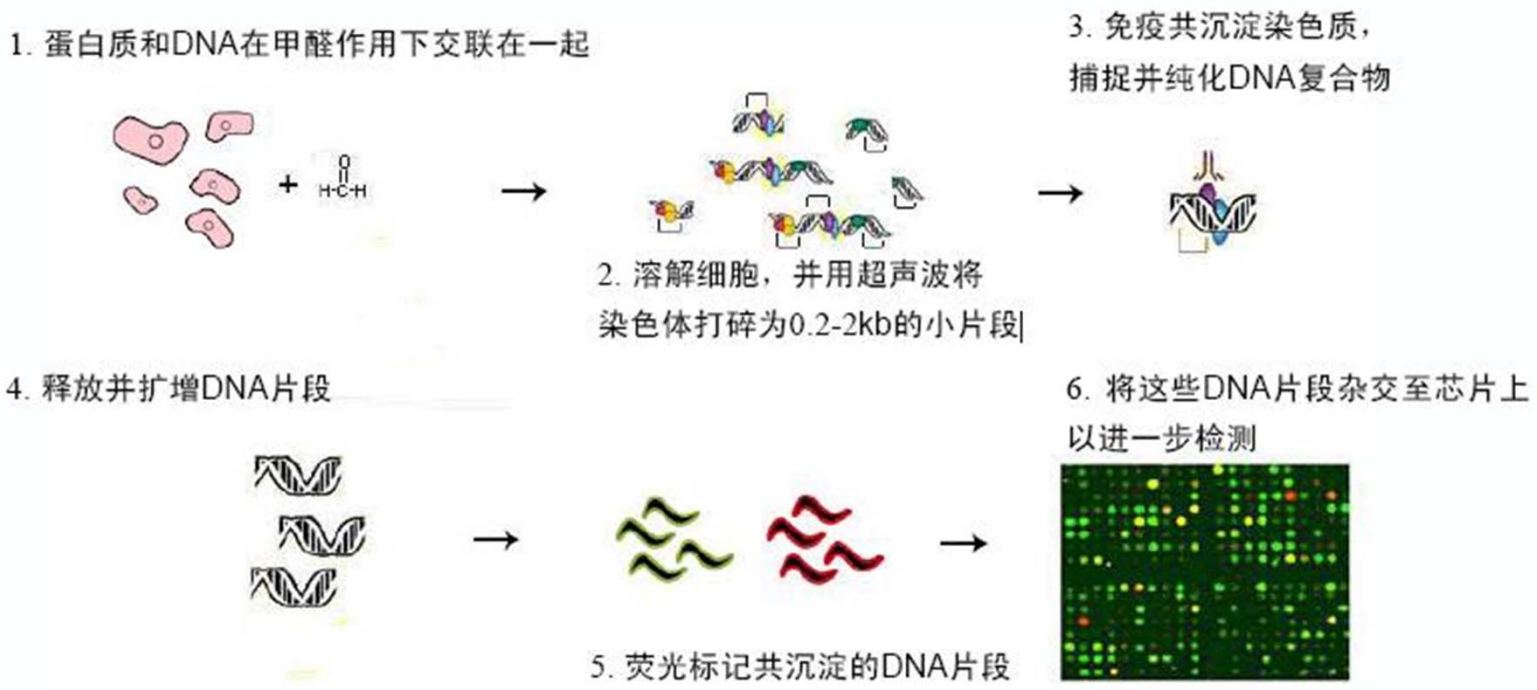
3. 集成大量的先验信息有助于缓解数据稀缺状况

4. 同时使得得到的调控网络更加精确。



Integrating prior

染色体免疫共沉淀技术（Chromatin Immunoprecipitation，ChIP）

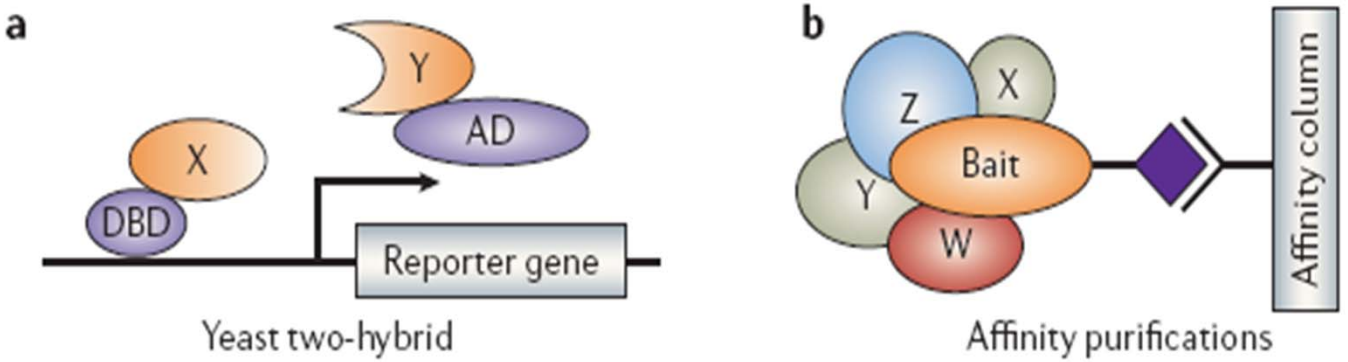


染色体免疫共沉淀在过去十年已经成为表观遗传信息研究的主要方法，确定转录因子及其结合位点

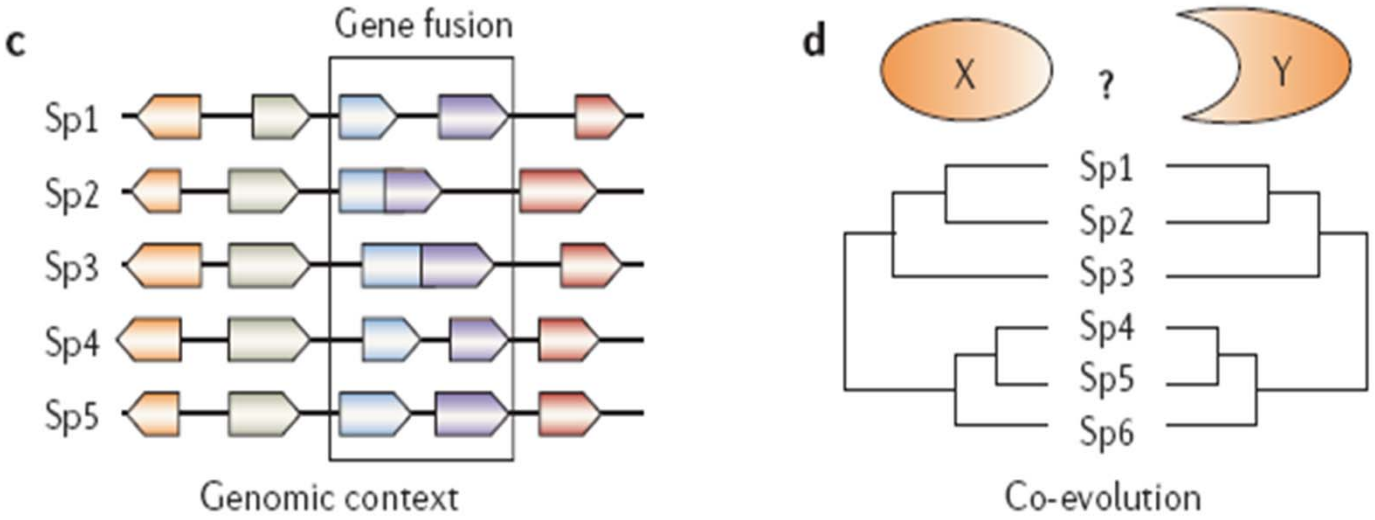
实验方法预测蛋白相互作用

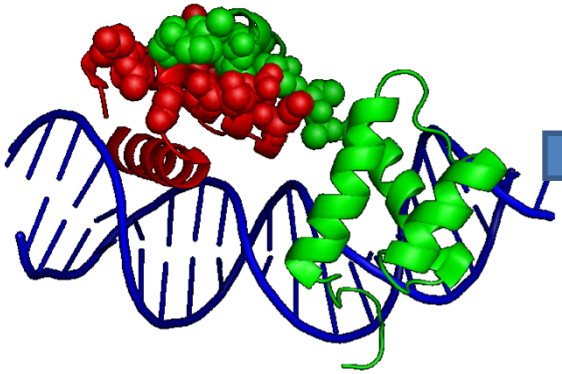
Box 1 | **Uncovering protein interactions**

Experimental methods



Computational methods





a

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Source binding sites

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

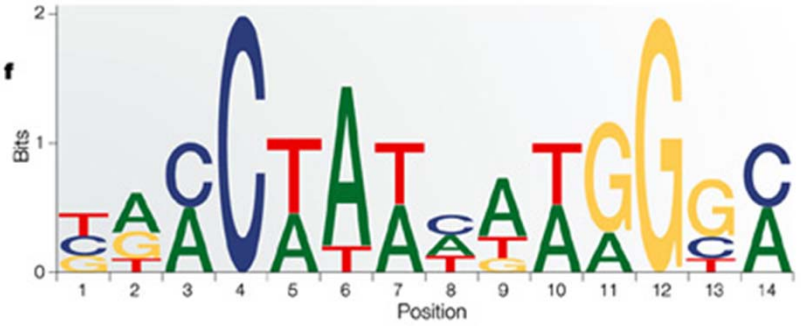
A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

e Site scoring

0.45	−0.66	0.79	1.68	0.45	−0.66	0.79	0.45	−0.66	0.79	0.00	1.68	−0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 78% of maximum

f



Nature Reviews | Genetics

先验信息分类

- **无向(Undirected):** 仅仅知道有无调控关系。例如蛋白质相互作用数据以及共表达数据
- **有向无符号(Directed and un-signed).** 知道有方向的调控关系，但是不知道是激活还是压制作用。例如ChIP-chip 数据和motif 出现数据
- **有向有符号(Directed and signed).** 知道有方向的调控关系，同时知道是激活还是压制作用，但是没有调控的强度数据，例如文献中记录的调控关系

集成先验信息的线性规划模型

有很多对网络结构推断有价值的的先验信息，例如从数据库或文献中得到的基因间调控数据，这些信息可通过添加线性规划的约束来提高所得到的聚合网络的精度。

$$\min_{Y^1,Y^2,\ldots,Y^N,L} \quad \sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{n+s}\omega_k \left| L_{ij} - L_{ij}^k \right| + \lambda \sum_{(i,j)\in\{(i,j)|K_{ij}=0 \text{ or } U_{ij}=0\}} \left| L_{ij} \right|$$

$$s.t. \quad L_{ij} > 0 \quad \text{if} \quad K_{ij} > 0 \quad i,j \in \{1,2,\ldots,n\}$$

$$L_{ij} < 0 \quad \text{if} \quad K_{ij} < 0 \quad i,j \in \{1,2,\ldots,n\}$$

$$L_{ij} = 0 \quad \text{if} \quad E_{ij} = 0 \quad i,j \in \{1,2,\ldots,n\}$$

硬约束：已知信息较为精确，希望在推断的网络中体现

软约束：噪声较大的信息，在推断的网络中出现与否取决于其他数据的相容性

GNMInfer (Gene Network reconstruction tool with Modular structure)

- Primary literature and information in databases for well-studied organisms such as E. coli and S. cerevisiae indicated the complex network takes network motifs and modules as its basic building block.
- Introducing the assumption is a cellular system is composed of locally interacting biological modules.
- Integrate the bottom-up and top-down reconstruction strategies.
- Initially perform a network modules identification. Then the modular gene regulatory network inferred from multiple microarray datasets To relieve the curse of dimension.
- To ensure sparse network in a structured way.

Top-down methodology

- Inferring a regulatory network without a priori knowledge
- TOP-DOWN APPROACH: *the architecture of the network is inferred (or reverse engineered) based on the observed response of the system to a series of experimental perturbations.*
- In engineering sciences: system identification

  1. typical use: large scale modeling from high throughput data (genomic/proteomic/metabolomic)

  2. main use: gene networks, any kind of complex network (metabolic, signalling pathways, protein activity, etc.)

Bottom-up methodology

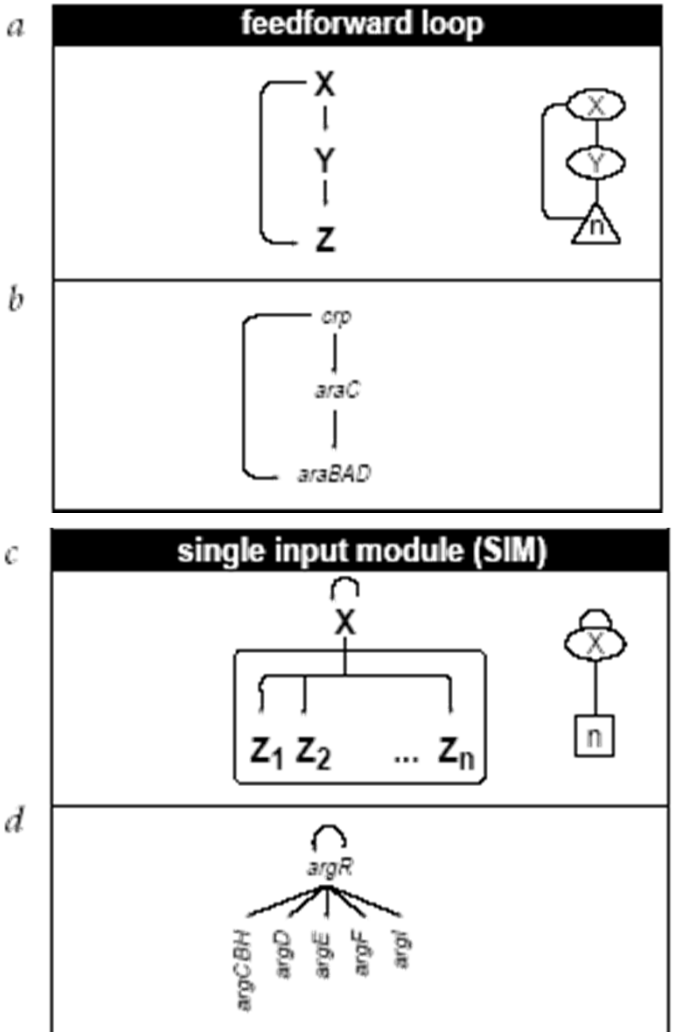
- Mathematical model was obtained from already available knowledge of the mechanisms of action/interaction between to or more components

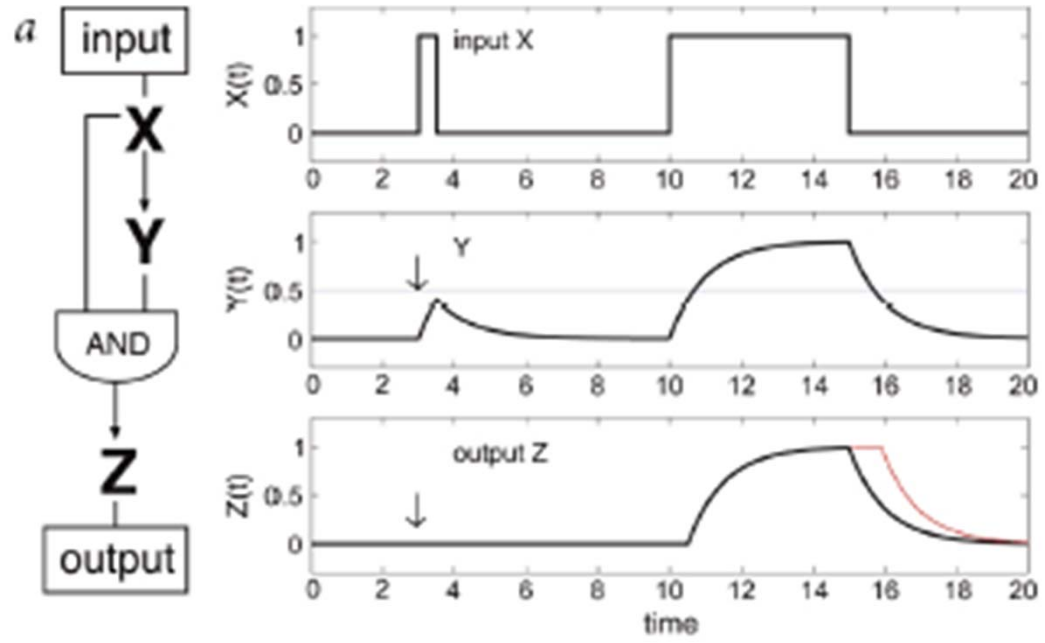
- BOTTOM-UP APPROACH: *model built from a priori biological information*

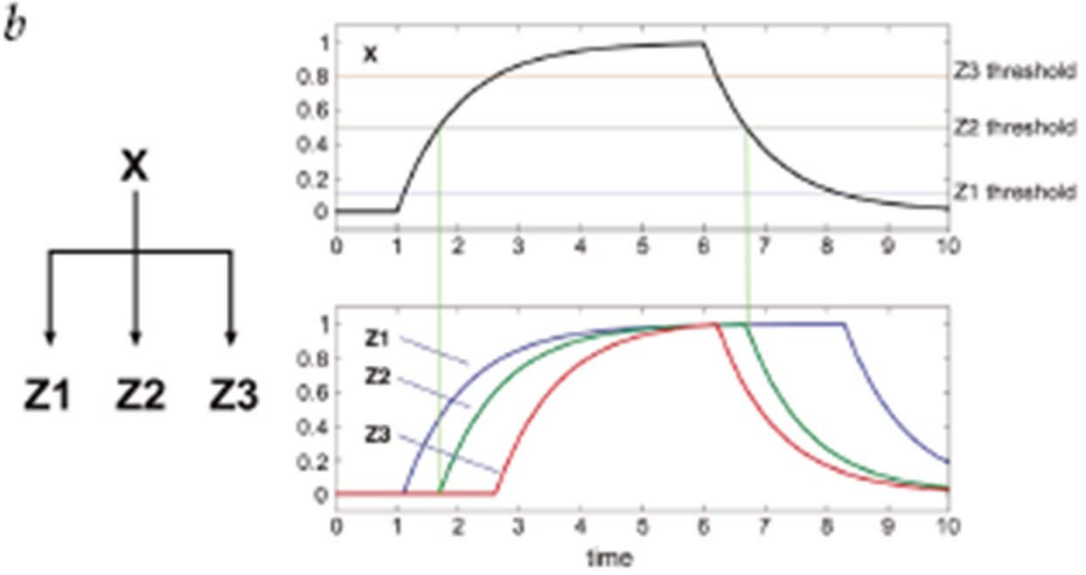
- Advantages:
  1. readily testable comparing simulation vs experiments
  2. allows to model known pathways
  3. allows to pass from qualitative to quantitative analysis

- Drawbacks
  1. can model only *known molecular processes*
  2. does not allow to *discover new pathways*
  3. less applicable to poorly characterized networks
  4. useful mainly for small/medium scale systems

Network motifs



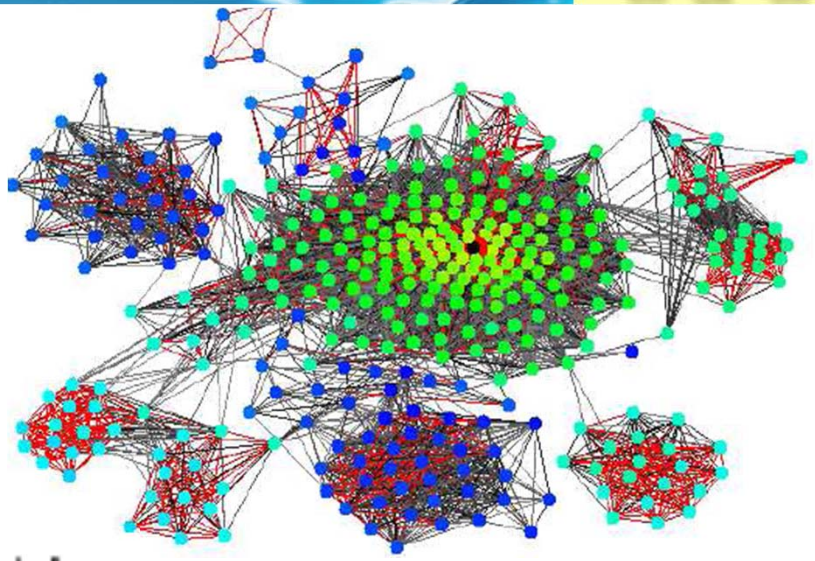


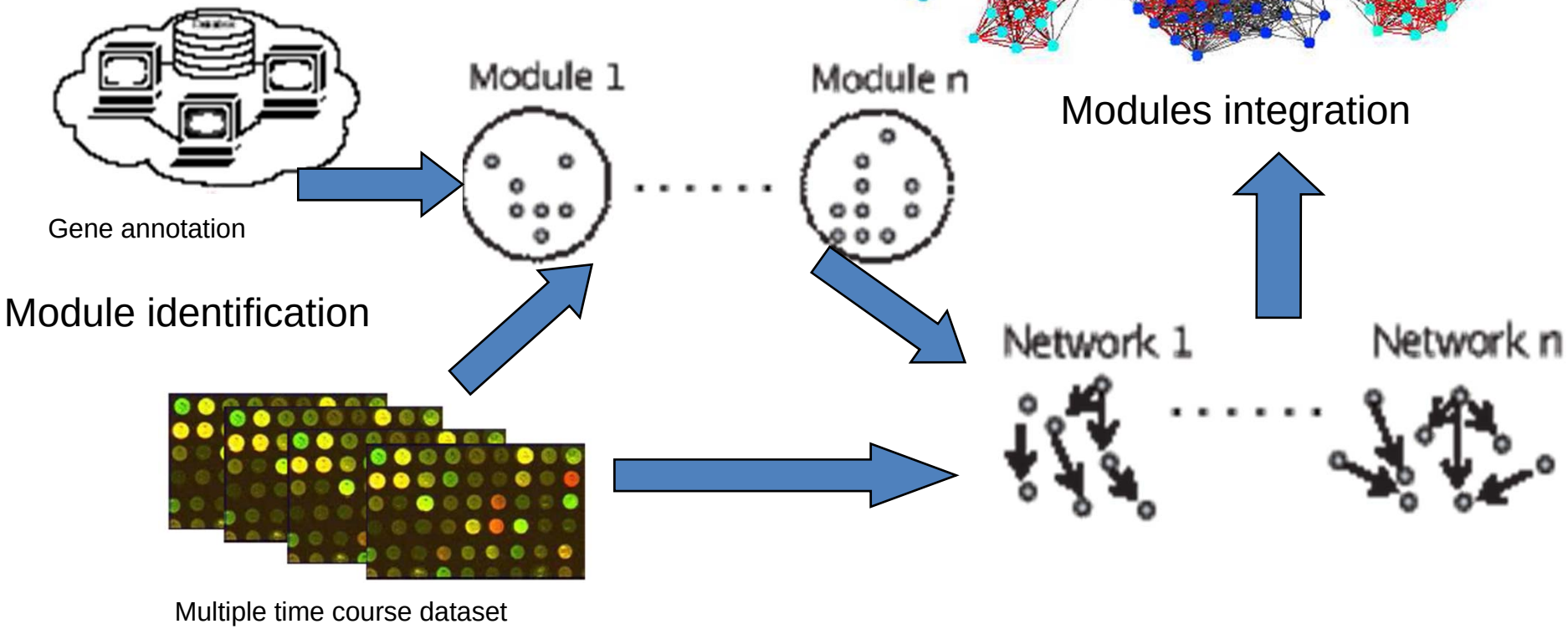


Network Modules

- Topological module: most of the genes are likely to be related to the genes in the same module rather than the genes in different modules. (Clustering on the expression data to find the co-regulations relationships)
- Functional module: most of the genes are likely to have similar function related to the genes in the same module rather than the genes in different modules. (Clustering the gene annotation data to find the similar function relationships)

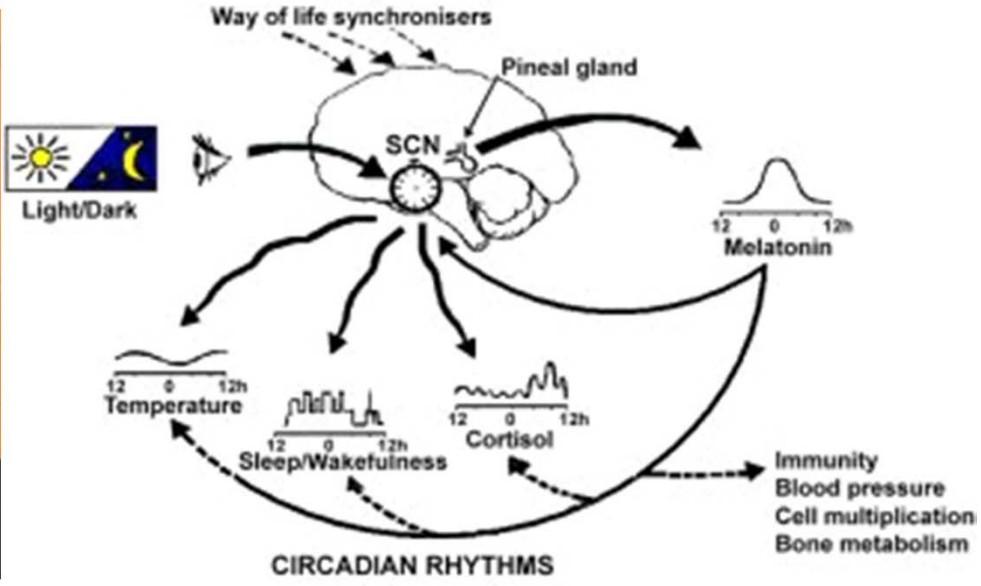
Flowchart of GNMInfer





An Example: Circadian rhythm

- Circadian rhythm is fundamentally important in physiological processes of mammals.



生命活动以24小时左右为周期的变动。又称近日节律。发光菌的发光，植物的光合作用，动物的摄食，躯体活动，睡眠和觉醒等行为显示昼夜节律。人体生理功能，学习与记忆能力、情绪、工作效率等也有明显的昼夜节律波动。

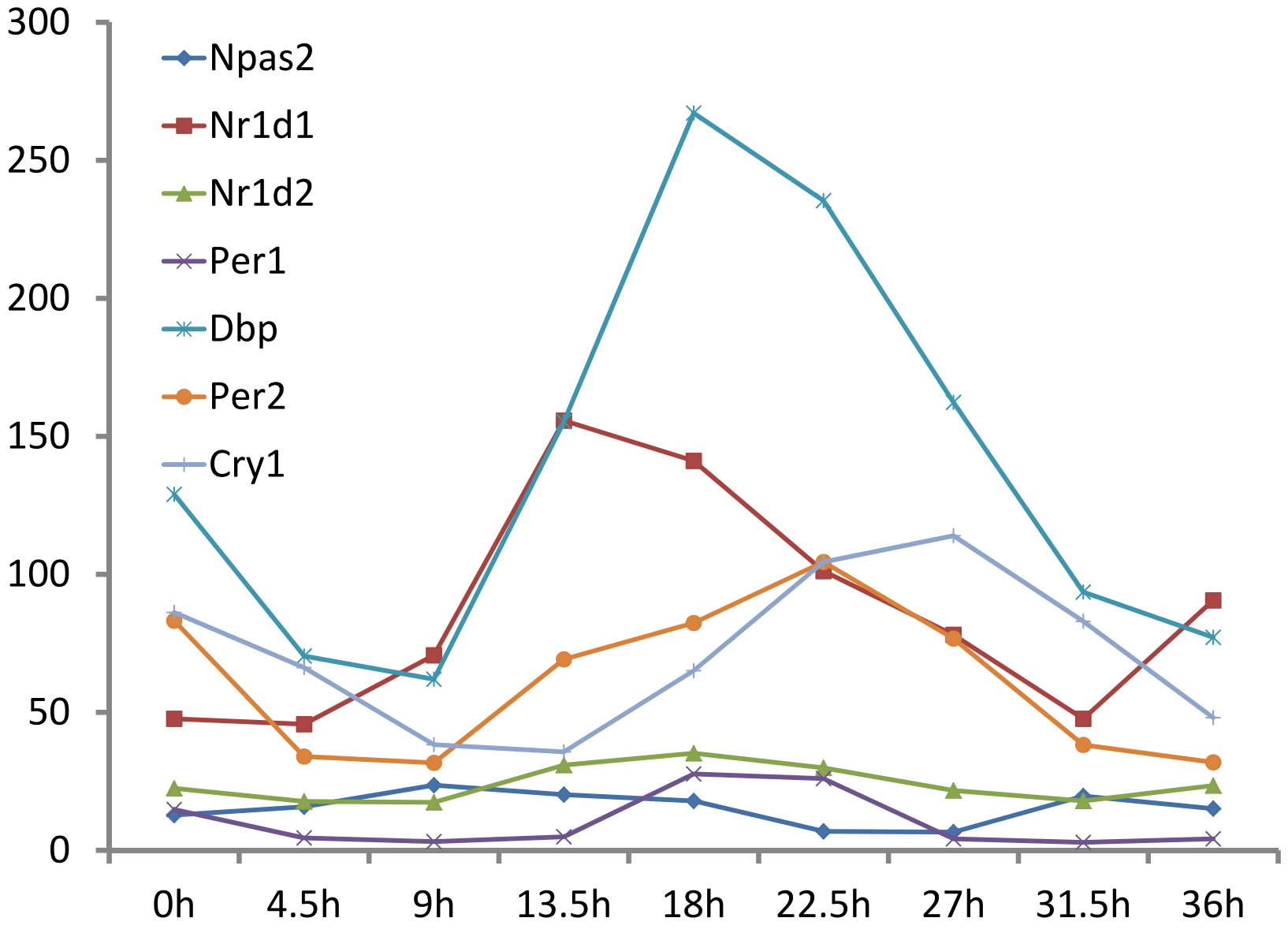
Why gene regulatory network

- The 20,000 dissociated neurons consisting of a pair of the mammalian uprachiasmatic nuclei (SCN) display autonomous rhythms in electrophysiological activities. This indicates that the oscillator mechanism resides within individual cells
- Recent observations revealed that a large number of genes undergo circadian oscillation in their expression levels.
- Furthermore, extensive studies have identified that a set of key circadian genes utilize the transcriptional-translational auto-regulatory loop to generate molecular oscillations of the “central clock”.

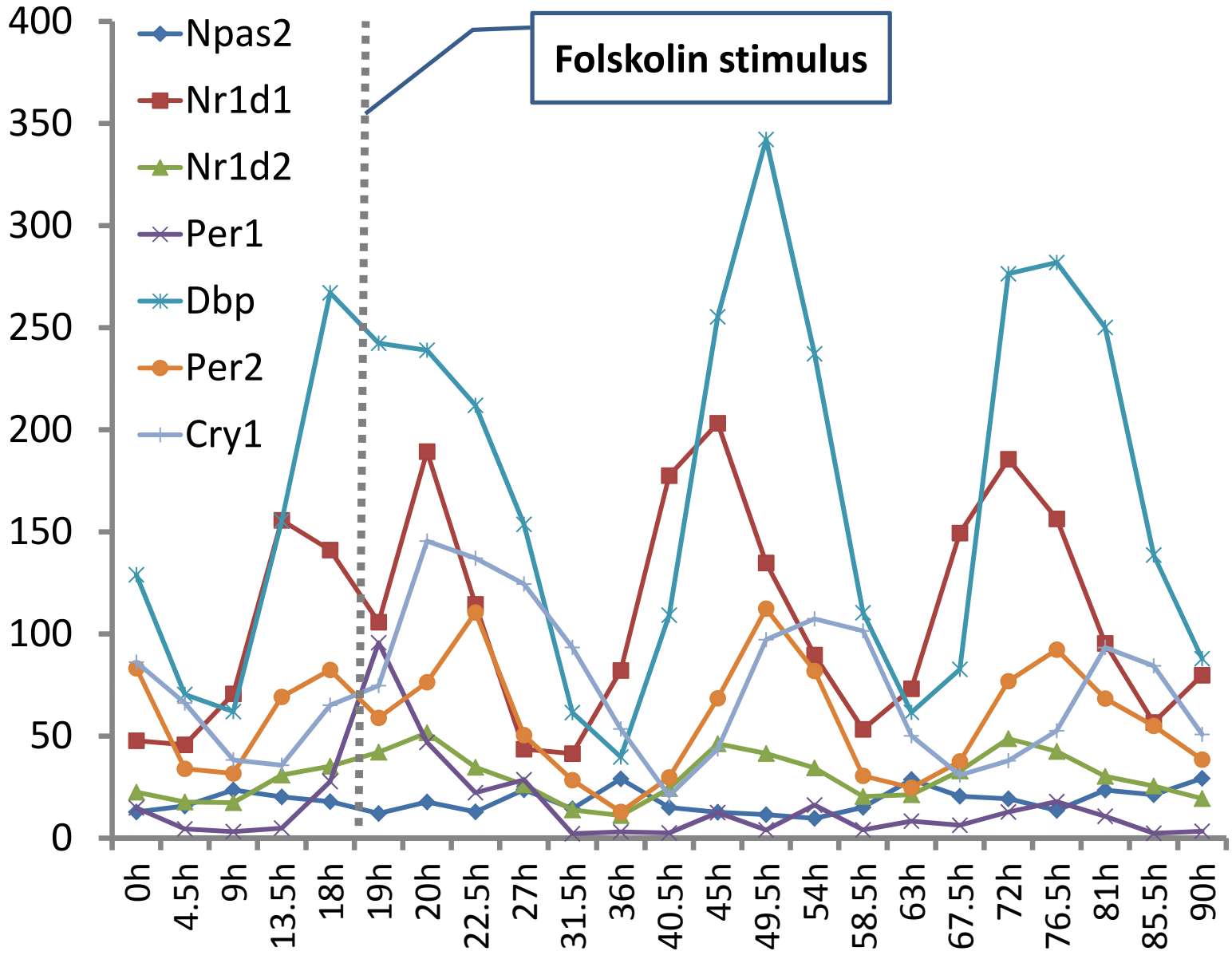
Gene expression data

- The laboratory cultured cells from SCN
- Perturbation: Forskolin stimuli can reset the clock of the cells by phase advance and phase delay.
- Four time-series microarray

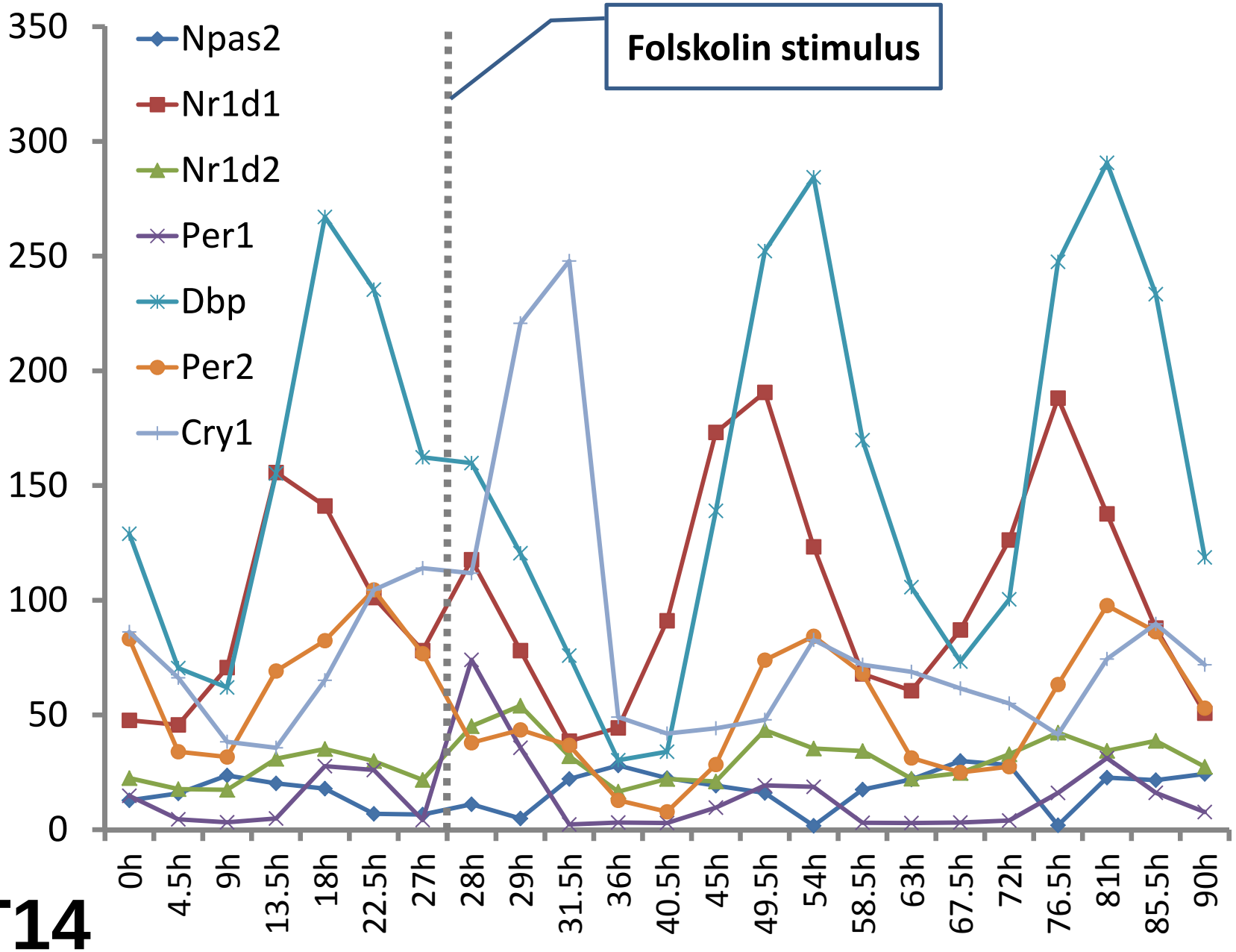
1. Control, 0-36 hour, **14** time points;
2. CT6, 0-90 hour, drug is applied at 18 hour, **16** time points;
3. CT14, 0-90 hour, drug is applied at 27 hour, **14** time points;
4. CT22,0-90 hour, drug is applied at 32 hour, **12** time points.


300
250
200
150
100
50
0
Npas2
Nr1d1
Nr1d2
Per1
Dbp
Per2
Cry1
0h
4.5h
9h
13.5h
18h
22.5h
27h
31.5h
36h

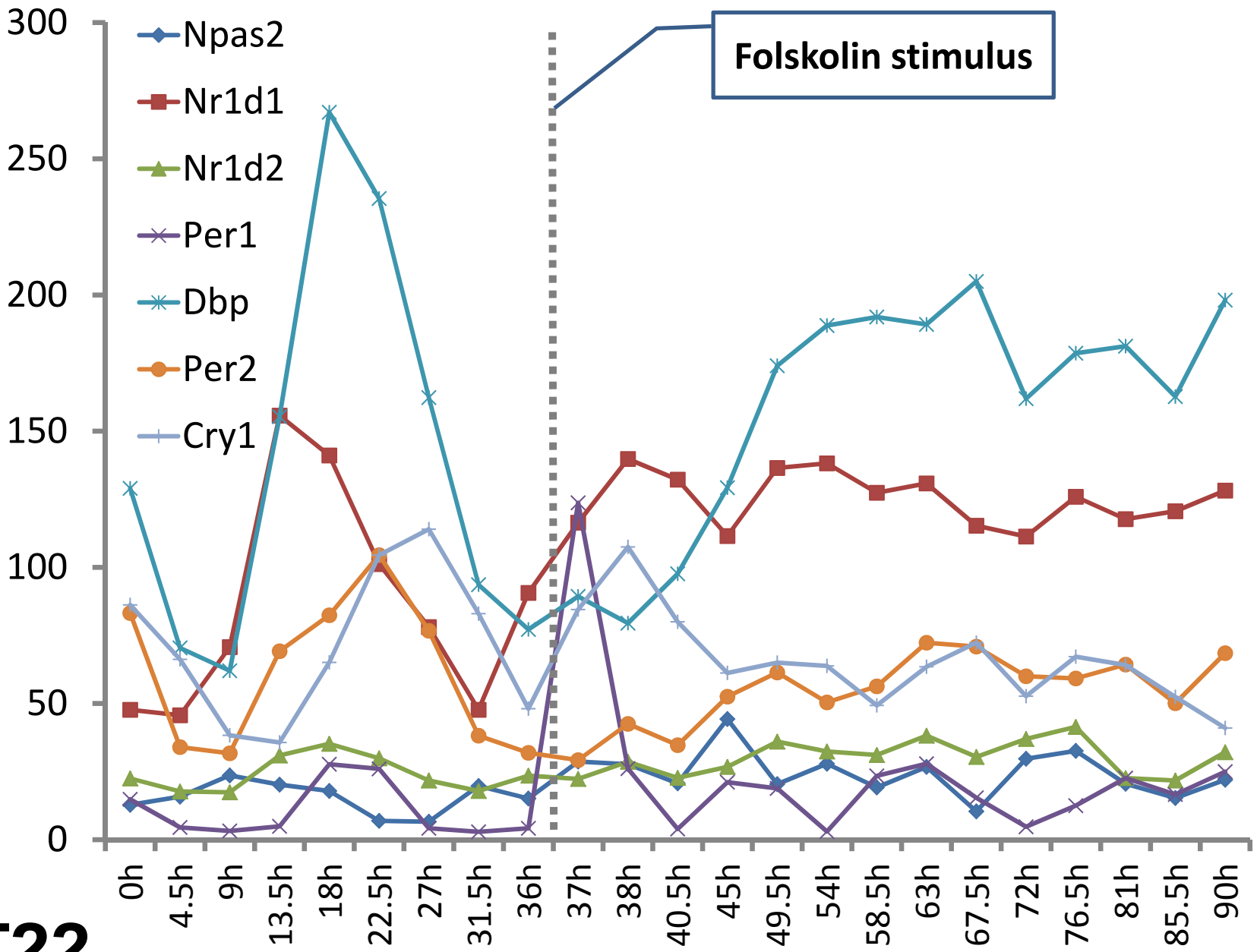
Control


400
350
300
250
200
150
100
50
0
Npas2
Nr1d1
Nr1d2
Per1
Dbp
Per2
Cry1
Folskolin stimulus
0h
4.5h
9h
13.5h
18h
19h
20h
22.5h
27h
31.5h
36h
40.5h
45h
49.5h
54h
58.5h
63h
67.5h
72h
76.5h
81h
85.5h
90h

CT6


Folskolin stimulus
Npas2
Nr1d1
Nr1d2
Per1
Dbp
Per2
Cry1
350
300
250
200
150
100
50
0
0h
4.5h
9h
13.5h
18h
22.5h
27h
28h
29h
31.5h
36h
40.5h
45h
49.5h
54h
58.5h
63h
67.5h
72h
76.5h
81h
85.5h
90h

CT14


Folskolin stimulus
Npas2
Nr1d1
Nr1d2
Per1
Dbp
Per2
Cry1
300
250
200
150
100
50
0
0h
4.5h
9h
13.5h
18h
22.5h
27h
31.5h
36h
37h
38h
40.5h
45h
49.5h
54h
58.5h
63h
67.5h
72h
76.5h
81h
85.5h
90h

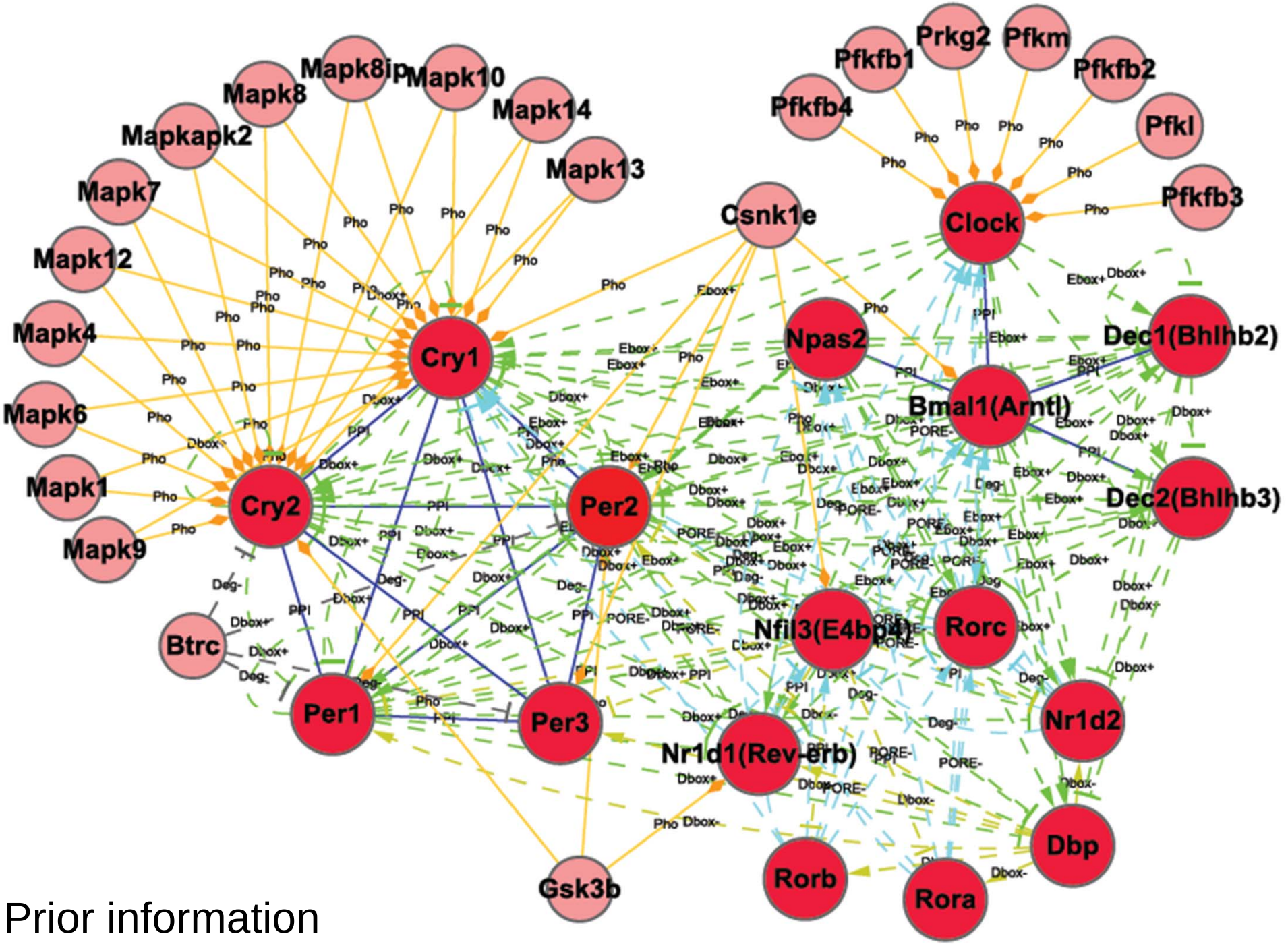
CT22

Candidate gene list

- **Key circadian genes: 18 well-studied clock genes**
- **Circadian-related genes: 22** genes having protein interactions and phosphorylations relationships with the 18 key circadian genes.
- **Oscillatory gene list:** 55 genes are identified to see whether typical oscillations exist or not in gene expression data.

Prior information

- **14 physical protein interactions**
- **40 phosphorylation interactions**
- ***Cis*-regulatory element:** 134 transcriptional regulatory interactions by linking the transcription factor with their target promoter region in the gene level
- **Protein-drug interaction:** the significantly induced and or repressed genes are identified as the potential target of the drug folskolin.



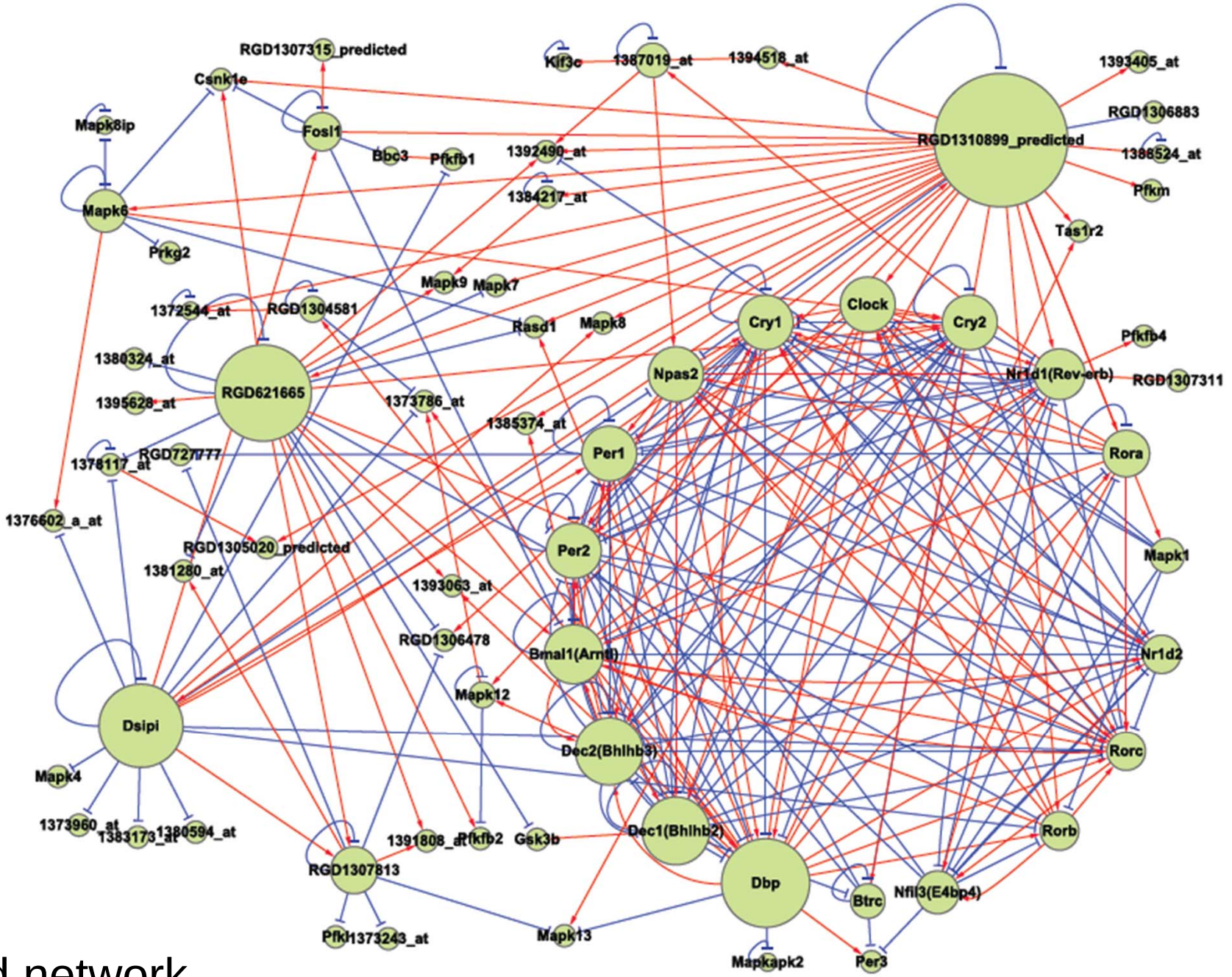
Prior information

Network inference

- 276 predicted regulatory relationships among 80 circadian related genes.
- 138 new regulatory relationships that are not in the prior information (73 activations and 65 repressions)

(a) brand new regulatory relationships

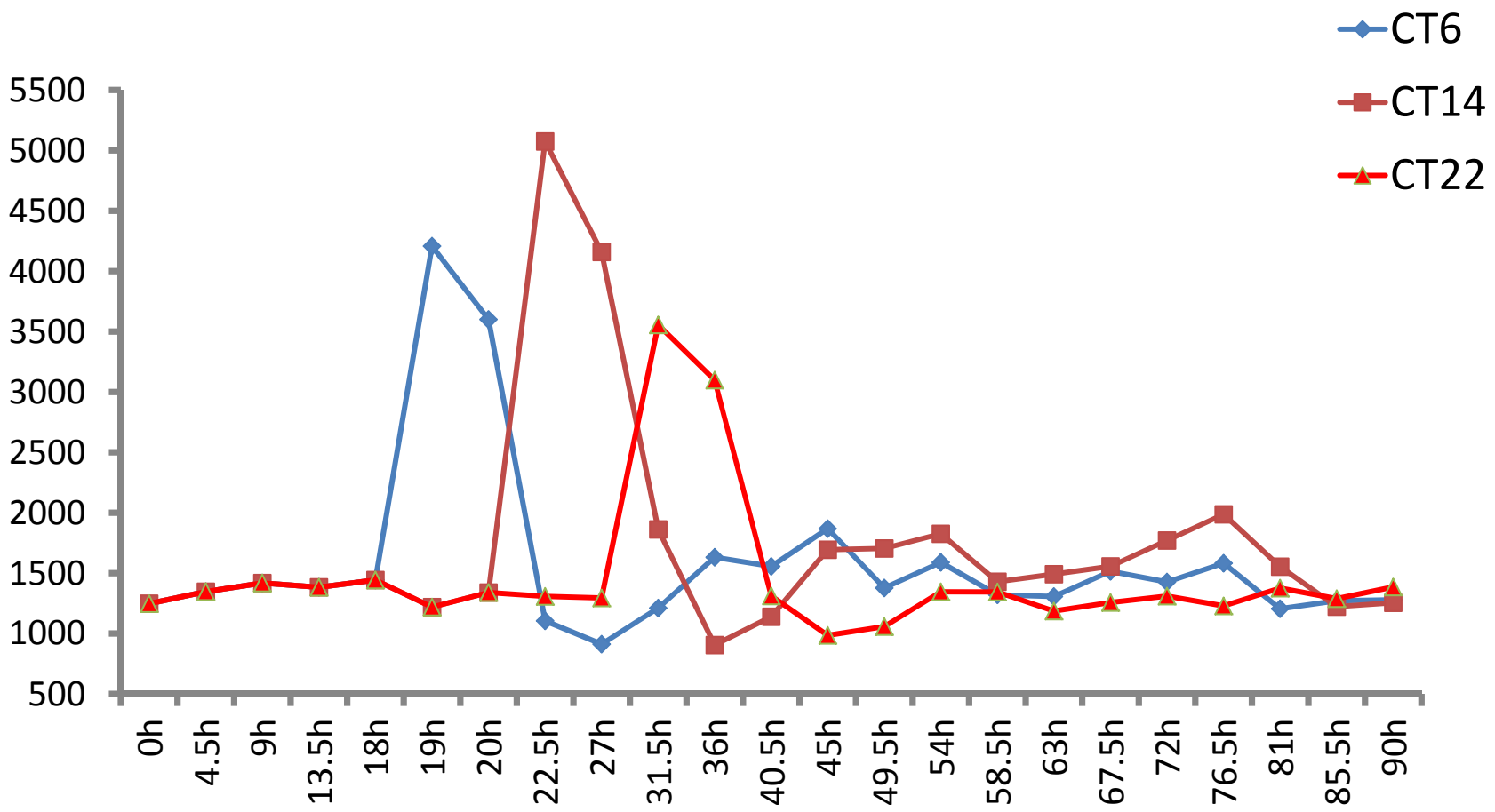
(b) signs and weights for those functional relationships in the prior information.



Inferred network

Four important hubs

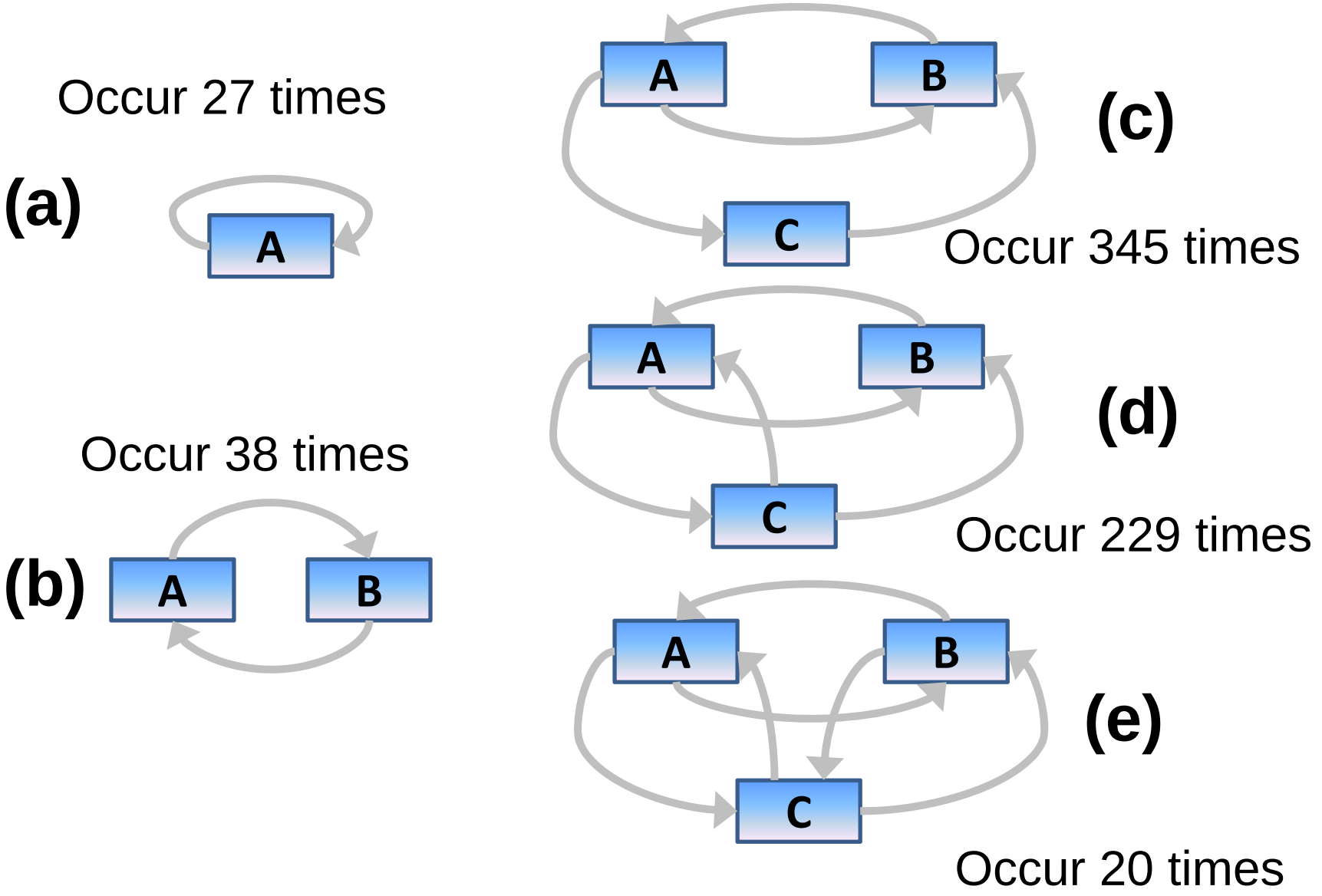
- *Dsipi* (regulate 17 target genes): A transcription factor protecting T-cells from IL2 deprivation-induced
- *RGD621665* (regulate 20 target genes): a regulator of G-protein signaling
- *RGD1307813* (regulate 8 target genes): related to endoplasmic reticulum,cell redox homeostasis, and protein folding.
- *RGD1310899_predicted* (regulate 29 target genes)



mRNA expression profile of hub gene *RGD621665 (Rgs2)*

Enriched motifs

- Transcription-translation feedback loops are important in driving circadian rhythm. For example, *Bmal1* and *Clock* proteins form a complex that positively regulates the transcription of *Per* and *Cry* family genes.
- z-score and p-value are used to assess the statistical significance of the certain motif in our predicted network against 1000 randomized networks



Enriched feedback motifs (p-value<1e-10)

An integrated gene regulatory network inference pipeline

Yong Wang, Katsuhisa Horimoto, and Luonan Chen

Academy of Mathematics and Systems Science, Chinese Academy of Mathematics, Beijing, 100190, China

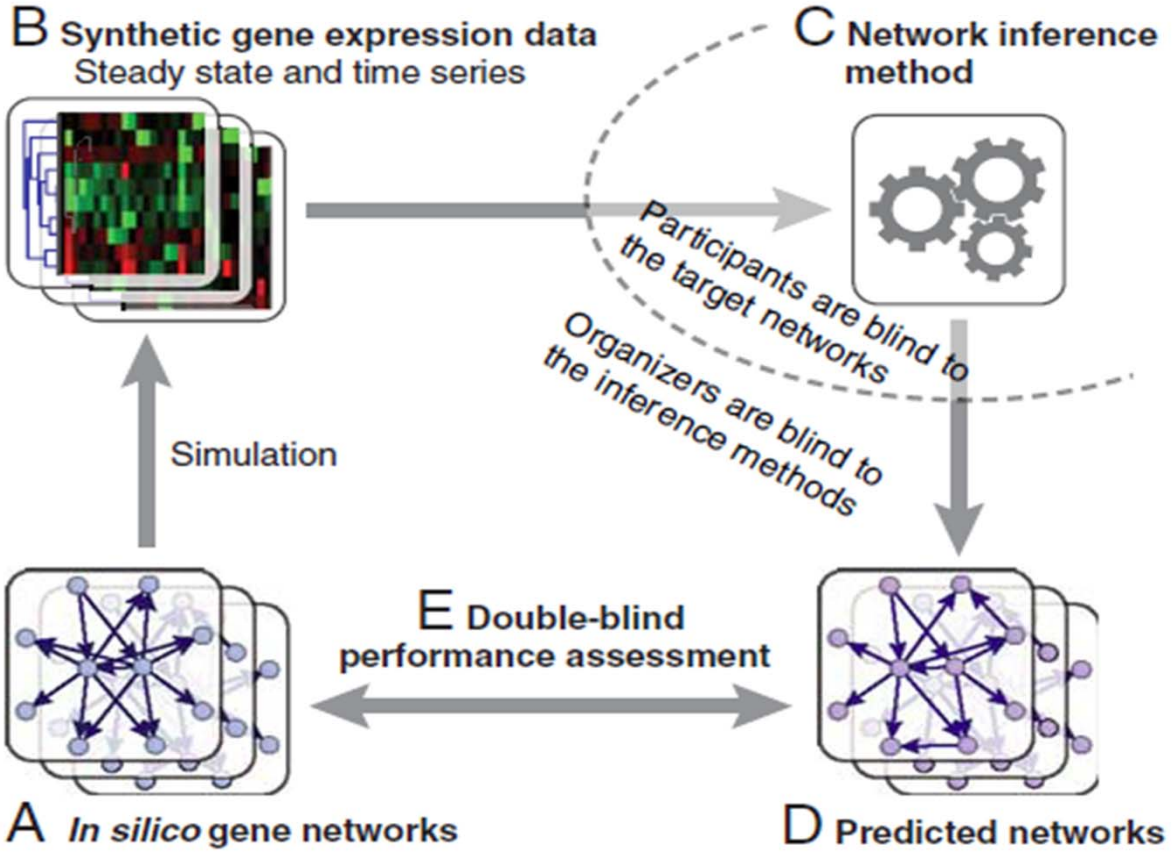
E-mail: y.wang@amss.ac.cn

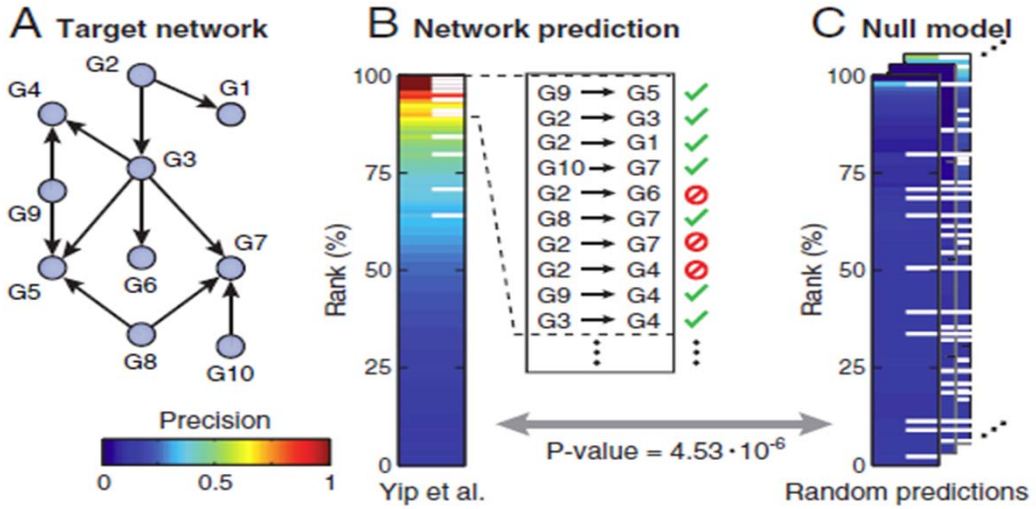
http://zhangroup.aporc.org

The 30th Chinese Control Conference, July 22-24, Yantai, China

DREAM

- Dialog on Reverse Engineering Assessment and Methods
- Annual workshop for evaluation of algorithms
- Last workshop had 40 teams participating
- Simulated mRNA expression profiles produced from an ODE model, including
  - All single-gene deletion mutants, grown the same way
  - Time course of mRNA expression after change in growth conditions


B Synthetic gene expression data
Steady state and time series
C Network inference method
Participants are blind to the target networks
Organizers are blind to the inference methods
Simulation
E Double-blind performance assessment
A In silico gene networks
D Predicted networks


A Target network
G2
G4
G1
G3
G9
G7
G5
G6
G8
G10
Precision
0
0.5
1
B Network prediction
100
75
50
25
0
Rank (%)
G9 → G5
G2 → G3
G2 → G1
G10 → G7
G2 → G6
G8 → G7
G2 → G7
G2 → G4
G9 → G4
G3 → G4
P-value = 4.53 · 10-6
Yip et al.
C Null model
100
75
50
25
0
Rank (%)
Random predictions

Follow-up analysis

- @article{madar2010dream3, title={{**DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator**}}, author={Madar, A. and Greenfield, A. and Vanden-Eijnden, E. and Bonneau, R.}, journal={PLoS ONE}, volume={5}, number={3}, pages={e9803}, year={2010} }
- @article{scheinine2009inferring, title={{**Inferring gene networks: dream or nightmare**?}}, author={Scheinine, A. and Mentzen, W.I. and Fotia, G. and Pieroni, E. and Maggio, F. and Mancosu, G. and de La Fuente, A.}, journal={Annals of the New York Academy of Sciences}, volume={1158}, number={The Challenges of Systems Biology Community Efforts to Harness Biological Complexity}, pages={287--301}, year={2009}, publisher={John Wiley \& Sons} }
- @article{marbach2010revealing, title={{**Revealing strengths and weaknesses of methods for gene network inference**}}, author={Marbach, D. and Prill, R.J. and Schaffter, T. and Mattiussi, C. and Floreano, D. and Stolovitzky, G.}, journal={Proceedings of the National Academy of Sciences}, volume={107}, number={14}, pages={6286}, year={2010}, publisher={National Acad Sciences} }
- @article{10.1371/journal.pone.0012912, author = {Pinna, Andrea AND Soranzo, Nicola AND de la Fuente, Alberto}, journal = {PLoS ONE}, publisher = {Public Library of Science}, title = {**From Knockouts to Networks: Establishing Direct Cause-Effect Relationships through Graph Analysis**}, year = {2010}, month = {10}, volume = {5}}

General conclusion

- Negative: reliable network inference from gene expression data remains an unsolved problem.

- Positive: the results indicate potential ways of network reconstruction improvements.

Detailed lessons learned

- The success is more related to the details of implementation than the choice of general methodology.
- Integration steady state data and time series data helps.
- Simpler method perform in general better than advanced, theoretically motivated approaches.
- Knock out data is useful

Motivation

- We maximally utilize the information in the limited gene expression data by categorizing the data into three types and developing three methods accordingly for information mining.
- We propose a path consistency algorithm based on conditional mutual information to differentiate the direct and indirect regulatory interactions.
- We integrate three methods into a pipeline by considering their complementarities and high accuracy can be expected.

Three types of data

- Type one is the steady-state gene expression profile of knock-out or knockdown experiments.
- Type two is the steady state gene expression profile after multi-factorial perturbations.
- Type three is the time-series data after multi-factorial perturbations.

Knockout data

- The idea is natural to identify if a gene x_i is a target of gene x_j by comparing the expression level of x_i when x_j is knocked out or knocked down to the wild-type expression of x_i.
- T-test: $T= (x_{ij}^{ko}- (x_1^{wt}+ x_2^{wt}+ \ldots+ x_N^{wt})/N)/\sigma$
- Fold change: $F= x_{ij}^{ko}/(x_1^{wt}+ x_2^{wt}+ \ldots+ x_N^{wt})/N)$
- Combine the two scores together

Steady state data

- Difficulty I: the non-linear relationships due to time-delay and other complicated factors.
  Strategy: a mutual information based framework
- Difficulty II: the causal relationships

  Strategy: path consistency algorithm based on conditional mutual information

Entropy

- Entropy (self-information)

$$H(p) = H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x)$$

  - *the amount of information* in a random variable
  - *average uncertainty* of a random variable
  - *the average length of the message* needed to transmit an outcome of that variable

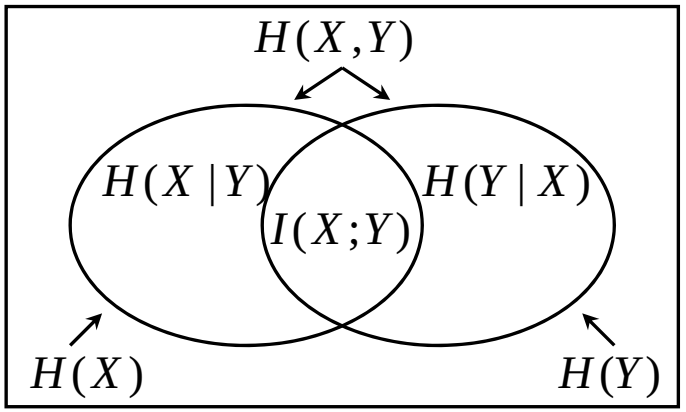
  - Properties
    - $H(X) \geq 0$ ($H(X) = 0$ providing no new information)

Mutual Information

- Mutual Information

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$



  - the reduction in uncertainty of one random variable due to knowing about another
  - the amount of information one random variable contains about another
  - measure of independence
    - two variables are independent $I(X;Y) = 0$
    - grows according to ...
      - the degree of dependence
      - the entropy of the variables

Conditional Mutual Information by Gaussian Kernel Estimator

- Assume N samples for Z_i with i=1,…,N

$$P(Z_i) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{(2\pi)^{n/2}\,|\mathrm{C}|^{n/2}} \exp\left(-\frac{1}{2}(Z_j - Z_i)^T C^{-1}(Z_j - Z_i)\right)$$

(j not equal i for N-1)

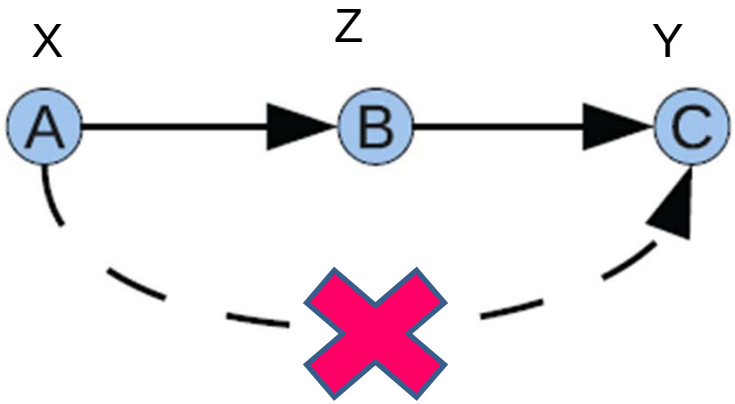
$$H(\mathrm{Z}) = -\frac{1}{N}\sum_{i=1}^{N} \ln(P(Z_i))$$

where Z_i is an n-dimensional vector of sample-i, and C is the covariance matrix of Z.

Then, we can estimate conditional mutual information based on this equation for H:

$$I(X, Y \mid Z) = H(X, Z) + H(Y, Z) - H(Z) - H(X, Y, Z)$$

Conditional mutual information



Z — XY	00	01	10	11
0	0.25	0	0	0
1	0	0.25	0.25	0.25

$$
\begin{aligned}
H(X) = & -p(x=0)log_2p(x=0) - p(x=1) \\
& log_2p(x=1) = 1\ bit, \\
H(Y) = & 1\ bit, H(Z) \approx 0.8113\ bit, \\
H(X,Y) = & -p(xy=00)log_2p(xy=00) - p(xy=01) \\
& log_2p(xy=01) \\
& -p(xy=10)log_2p(xy=10) - p(xy=11) \\
& log_2p(xy=11) \\
= & 2\ bits, \\
H(X,Z) = & 1.5\ bits, H(Y,Z) = 1.5\ bits, \\
H(X,Y,Z) = & -p(xyz=000)log_2p(xyz=000) \\
& -p(xyz=001)log_2p(xyz=001) \\
& -p(xyz=010)log_2p(xyz=010) \\
& -p(xyz=011)log_2p(xyz=011) \\
& -p(xyz=100)log_2p(xyz=100) \\
& -p(xyz=101)log_2p(xyz=101) \\
& -p(xyz=110)log_2p(xyz=110) \\
& -p(xyz=111)log_2p(xyz=111) \\
= & 2\ bits, \\
I(X;Y) = & H(X) + H(Y) - H(X,Y) = 0\ bit, \\
I(X;Y|Z) = & H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z) \\
\approx & \ 0.1887\ bit.
\end{aligned}
$$

Path Consistency Algorithm

- **Step 1.** Find a complete undirected subgraph (clique) with m nodes.
- **Step 2.** Calculate the zeroth-order conditional mutual information (for example mutual information of gene X and Y) and delete the edges that are independent.
- **Step 3.** Calculate the first-order conditional mutual information (for example mutual information of gene X and Y conditioning Z) and delete the edges that are independent.
- **Step 4.** Calculate the higher order conditional mutual information and terminate when there is no edges can be deleted.

Time series data

- We use the ordinary differential equation model to capture the dynamic relationship among genes

$$\frac{d\mathbf{X}}{dt} = J\mathbf{X} + P\mathbf{C}$$

$$\min_{Y^1, Y^2, \ldots, Y^N, L} \quad \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{n+s} \omega_k \mid L_{ij} - L_{ij}^k \mid + \lambda \sum_{(i,j) \in \{(i,j) \mid K_{ij}=0 \text{ or } U_{ij}=0\}} \mid L_{ij} \mid$$

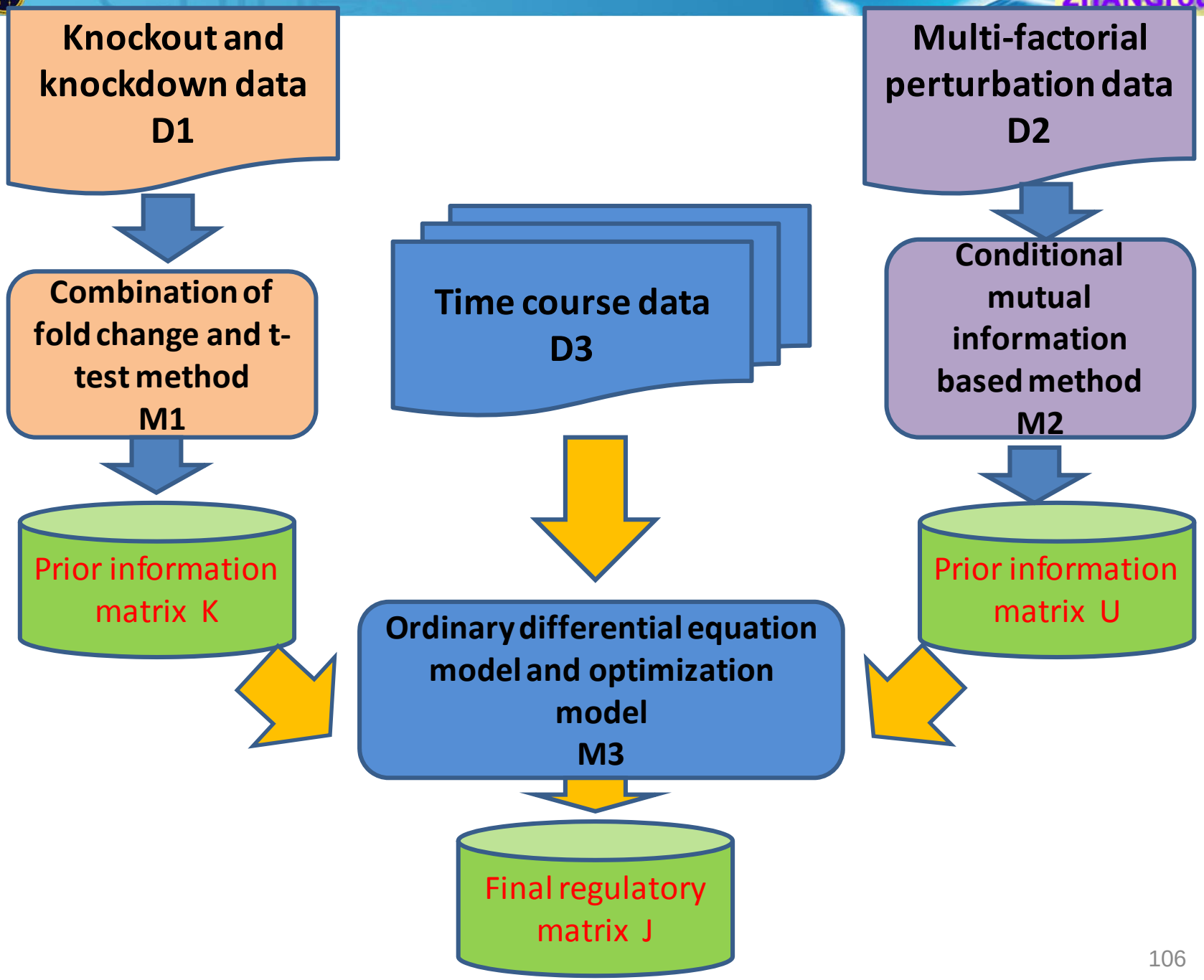
$$s.t. \quad L_{ij} > 0 \quad \text{if} \quad K_{ij} > 0 \quad i, j \in \{1,2,\ldots,n\}$$

$$L_{ij} < 0 \quad \text{if} \quad K_{ij} < 0 \quad i, j \in \{1,2,\ldots,n\}$$

$$L_{ij} = 0 \quad \text{if} \quad E_{ij} = 0 \quad i, j \in \{1,2,\ldots,n\}$$

Network inference pipeline

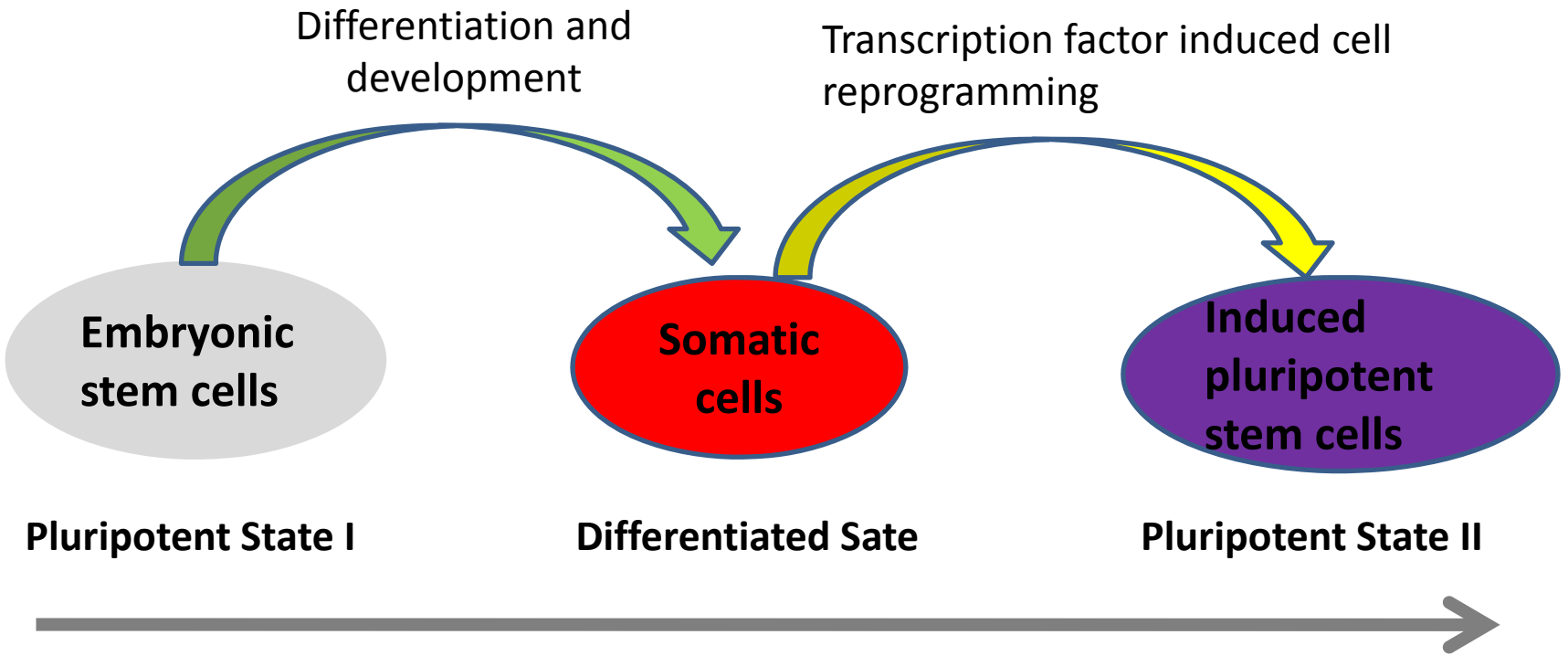
- D_1 (knockout and knockdown data)
- D_2 (steady state data)
- D_3 (time series data).
- M_1 (combination of fold change and t-test)
- M_2 (path consistency algorithm based on conditional mutual information)
- M_3 (ODE modeling of time series data).
- **$P(D|M)=P(D_1,D_2,D_3|M_1,M_2,M_3)=P(D_1|M_1,M_2,M_3)P(D_2|D_1,M_1,M_2,M_3)P(D_3|D_1,D_2,M_1,M_2,M_3)=P(D_1|M_1)P(D_2|M_2)\ P(D_3|\ D_1,D_2,M_1,M_2,M_3)$**


Knockout and knockdown data D1
Multi-factorial perturbation data D2
Combination of fold change and t-test method M1
Time course data D3
Conditional mutual information based method M2
Prior information matrix K
Prior information matrix U
Ordinary differential equation model and optimization model M3
Final regulatory matrix J

Advantages

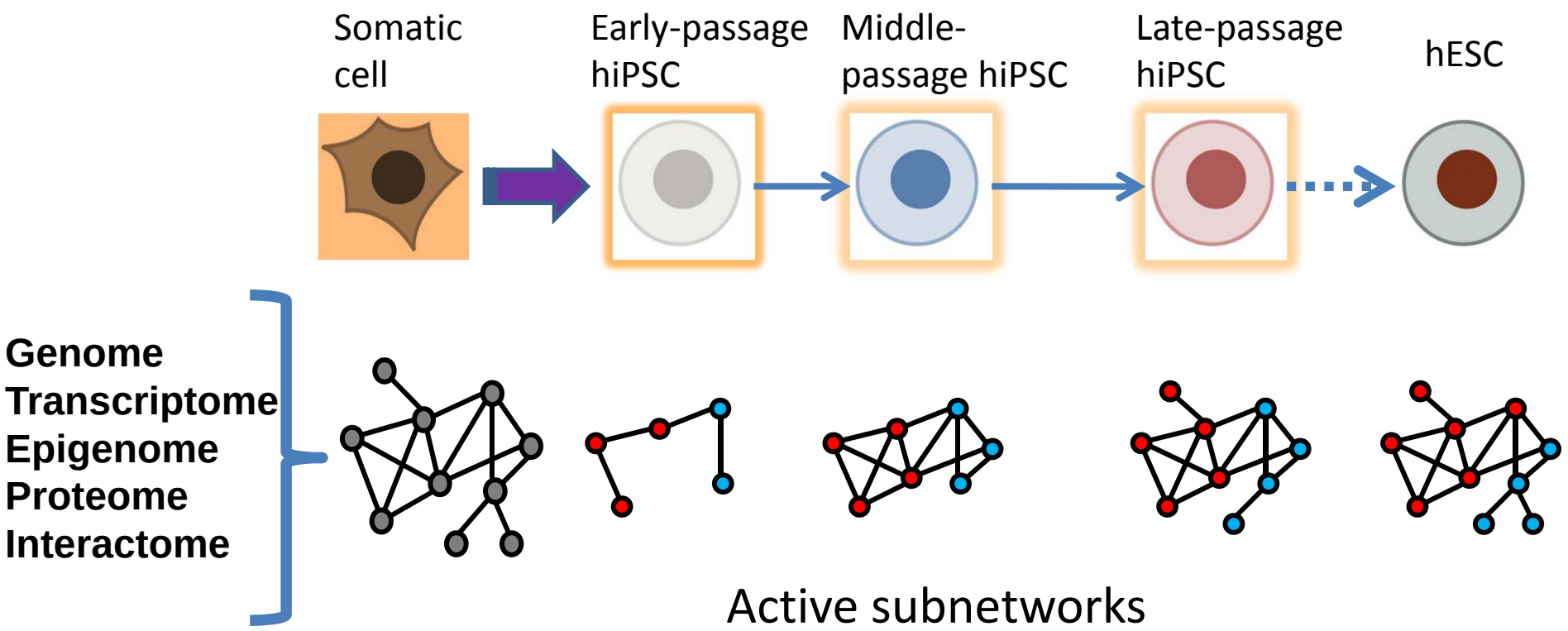
- Conditional mutual information is useful to reveal the hidden nonlinear relationships among genes.
- ODE model with prior information can predict the combinatorial regulations.
- Path consistency algorithm can remove the indirect regulations.
- Maximally utilizing information in the available data, emphasizing the knock-out and knock-down data, and differentiating the direct and indirect regulatory interactions.

Application: GRN for cell reprogramming



Question 1： Is the pluripotent state I identical with pluripotent state II ?
Question 2： How to use the high-throughput data to standardize iPSCs?
Question 3： What's the regulatory mechanism underlying cell reprogramming?

Network study for cell reprogramming



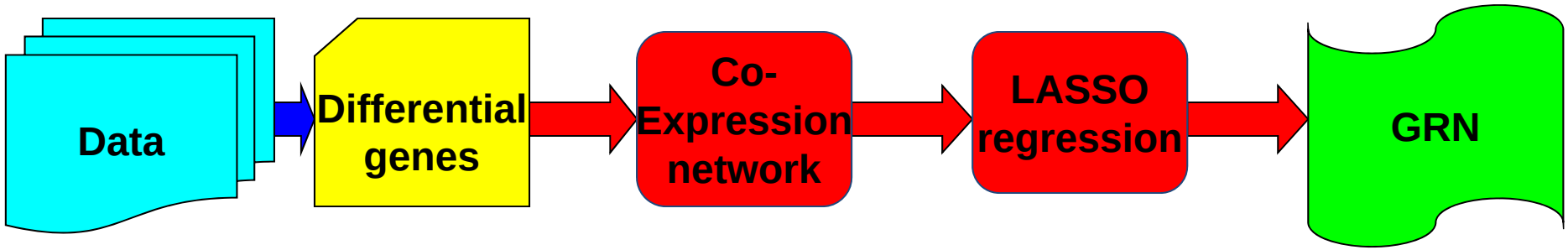
Public Time course data

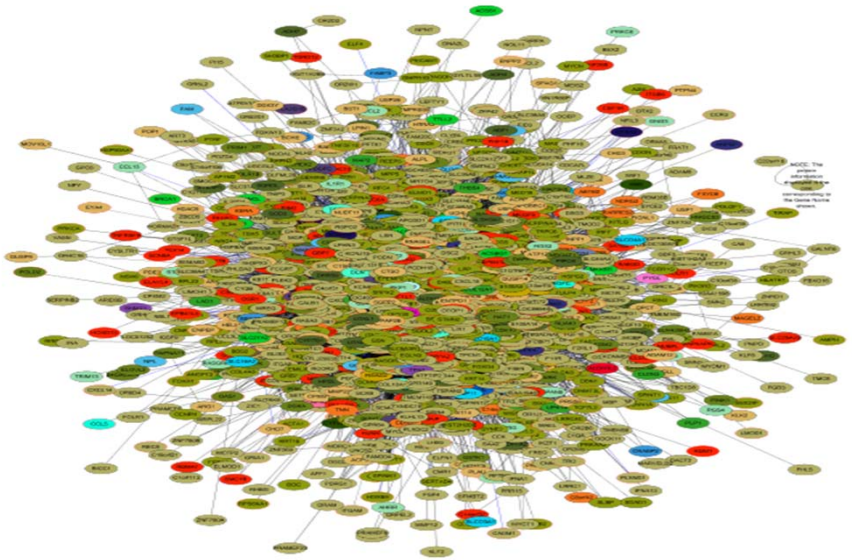
Data: The expression data is from [1] and measure throughout reprogramming of MEF to iPSC. Total RNA was extracted at day 0 (no Dox), day 2, 5, 8, 11, 16 and 21 (with Dox) and day 30 (Dox-independent secondary iPS). Therefore the data we used for network inference is time series data of 13,877 genes at 8 time points.

Observations: Temporal analysis of this time course data already revealed that reprogramming is a multi-step process that is characterized by initiation, maturation, and stabilization phases.

[1] Payman Samavarchi-Tehrani, et al., Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming, Cell Stem Cell, Vol. 7, No. 1, 2010, Pages 64-77

Preliminary result





Zhanna et al. CCC2012, 2012

Take home message

- Network study enables a system-wide overview on the gene regulation in mechanism of circadian rhythm.

- Data integration strategy improves the reliability of the inferred gene regulatory network.